MODERN COMPUTING METHODS

SECOND EDITION

PHILOSOPHICAL LIBRARY
NEW YORK

Published 1961
by Philosophical Library, Inc.
15 East 40th Street, New York 16, N.Y.

Printed in England for Philosophical Library
by John Wright & Sons Ltd., at the Stonebridge Press, Bristol

CONTENTS

Introduction . . . . *page* v

1. Linear Equations and Matrices: Direct Methods . . . 1

2. Linear Equations and Matrices: Direct Methods on Automatic Computers 13

3. Latent Roots and Vectors of Matrices 22

4. Linear Equations and Matrices: Iterative Methods . . 34

5. Linear Equations and Matrices: Error Analysis . . . 41

6. Zeros of Polynomials 53

7. Finite-difference Methods 62

8. Chebyshev Series 71

9. Ordinary Differential Equations: Initial-value Problems . . 80

10. Ordinary Differential Equations: Boundary-value Problems . 93

11. Hyperbolic Partial Differential Equations 101

12. Parabolic and Elliptic Partial Differential Equations . . 112

13. Evaluation of Limits; Use of Recurrence Relations . . 122

14. Evaluation of Integrals 130

15. Tabulation of Mathematical Functions 137

Bibliography 145

Index 167

INTRODUCTION

The first edition of Modern Computing Methods was based on lectures delivered by various members of the staff of Mathematics Division, N.P.L., as part of a vacation course on 'Computers for Electrical Engineering Problems', organized by the Electrical Engineering Department of the Imperial College of Science and Technology, and attended by representatives of industrial firms. The course was designed to teach the basic principles of the use of analogue machines, high-speed digital computers, and the techniques of numerical mathematics involved in the solution of problems in electrical engineering.

Numerical methods are required in all branches of science, and the techniques are generally independent of the source of the problem. For example, the same type of differential equation may represent a problem in physiology as well as a problem in electrical engineering. The opportunity was therefore taken to present, as one of the N.P.L. series of *Notes on Applied Science*, suitably edited versions of those lectures contributed to the course by members of Mathematics Division.

The success of this first edition has encouraged the authors to undertake a complete revision, in the course of which the booklet has been very largely rewritten. The object has been to bring the material up to date, particularly with regard to methods suitable for automatic computation, and the principal changes are as follows. The chapter on *Relaxation Methods* has been replaced by one on *Linear Equations and Matrices: Iterative Methods*, while the chapter in the original edition headed *Computation of Mathematical Functions* has been expanded into two chapters entitled *Evaluation of Limits; Use of Recurrence Relations* and *Evaluation of Integrals*. Most of the other chapters have had new material added, some being completely rewritten, and the order of the chapters has been changed. In addition, the chapters on *Linear Equations and Matrices: Error Analysis* and on *Chebyshev Series* are new.

The authors make no apology for the varying level of treatment of the different topics; this is inevitable if the account is to be kept within a comparatively small compass. Some of the new material is given in greater detail than classical material already well catered for in available text-books. It is hoped that the resulting booklet will prove useful both as a working manual for those engaged in computational work and as a basis for courses in numerical analysis in universities and technical colleges.

The first edition was written by L. Fox, E. T. Goodwin, J. G, L. Michel, F. W. J. Olver and J. H. Wilkinson. The present edition has been prepared by C. W. Clenshaw, E. T. Goodwin, D. W. Martin, G. F. Miller, F. W. J. Olver and J. H. Wilkinson, all members of the staff of Mathematics Division, N.P.L. Valuable criticisms and suggestions have also been made by L. Fox, now Director of the Oxford University Computing Laboratory, and many other members of Mathematics Division, particularly E. L. Albasiny, J. G. Hayes and T. Vickers. Mrs. I. Goode has collated the material, prepared the printer's copy and helped to see the work through the press.

E. T. GOODWIN
Superintendent
Mathematics Division
National Physical Laboratory

May 1960

1

LINEAR EQUATIONS AND MATRICES: DIRECT METHODS

DEFINITIONS AND PROPERTIES

1. A general set of n linear simultaneous algebraic equations in n unknowns $x_1, x_2, \ldots, x_n$ can be written in the form

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \;= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \;= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \;= b_n. \end{array}\right\} \tag{1}$$

The coefficients a_{rs} form a *square matrix of order n*,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \tag{2}$$

which is to be considered simply as an array of numbers, and the column of constants b_r similarly forms a *column matrix* or *vector* $\mathbf{b}$. The unknowns x_r form a vector $\mathbf{x}$. Equations (1) can then be written in the shortened form

$$\mathbf{Ax} = \mathbf{b}. \tag{3}$$

The equality sign means that each element of the product vector $\mathbf{Ax}$ is equal to the corresponding element of the vector $\mathbf{b}$, and the left of (1) gives the rule for *premultiplication* of a vector by a matrix.

2. The solution of (1) can be written in general terms as

$$\left.\begin{array}{l} x_1 = \alpha_{11}b_1 + \alpha_{12}b_2 + \ldots + \alpha_{1n}b_n, \\ x_2 = \alpha_{21}b_1 + \alpha_{22}b_2 + \ldots + \alpha_{2n}b_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = \alpha_{n1}b_1 + \alpha_{n2}b_2 + \ldots + \alpha_{nn}b_n, \end{array}\right\} \tag{4}$$

or in matrix notation as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \tag{5}$$

where $\mathbf{A}^{-1}$ has the same form as (2) with a replaced by α. The matrix $\mathbf{A}^{-1}$ is called the *inverse* or *reciprocal* of the matrix **A**.

The elements α_{rs} of $\mathbf{A}^{-1}$ depend only on the elements of **A**. It is clear from (4) that a knowledge of the α_{rs} would enable the solutions of (1) to be obtained with relative ease for any set of constants b_r, but the determination of the α_{rs} is not trivial. One method of obtaining them is suggested by equations (4): if in these equations we write

$$b_1 = 1, \quad b_2 = b_3 = \ldots = b_n = 0,$$

we obtain the elements of the first column of the inverse. It follows that the various columns of $\mathbf{A}^{-1}$ can be found in succession by solving equations (1) with the right-hand sides replaced by successive columns of the matrix

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & 0 & \ldots & . & 0 \\ 0 & 1 & 0 & \ldots & . & 0 \\ 0 & 0 & 1 & \ldots & . & 0 \\ . & . & . & \ldots & . & . \\ . & . & . & \ldots & . & . \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix}. \tag{6}$$

This is the *unit matrix* of order n or *identity matrix*, so called in virtue of the relation

$$\mathbf{I}_n \mathbf{x} = \mathbf{x} \tag{7}$$

for any vector **x**. The suffix n is usually omitted when there is no possible ambiguity.

3. The main properties of matrices required in practice are those of addition, multiplication, and transposition.

Matrices can be added only when of the same order, and if **B** is the matrix (2) in which a is replaced by b, then

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} a_{11}+b_{11}, & a_{12}+b_{12}, & \ldots, & a_{1n}+b_{1n} \\ a_{21}+b_{21}, & a_{22}+b_{22}, & \ldots, & a_{2n}+b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1}+b_{n1}, & a_{n2}+b_{n2}, & \ldots, & a_{nn}+b_{nn} \end{bmatrix}. \tag{8}$$

If a matrix **A** is multiplied by a number k, the resulting matrix has elements ka_{rs}, that is, every element is multiplied by k.

Square matrices of the same order can be multiplied, to give

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11}+a_{12}b_{21}+a_{13}b_{31}+\ldots, & a_{11}b_{12}+a_{12}b_{22}+a_{13}b_{32}+\ldots, & \ldots \\ a_{21}b_{11}+a_{22}b_{21}+a_{23}b_{31}+\ldots, & a_{21}b_{12}+a_{22}b_{22}+a_{23}b_{32}+\ldots, & \ldots \\ \ldots & \ldots & \ldots \end{bmatrix}. \tag{9}$$

If we use the notation $r_r(\mathbf{A}), c_r(\mathbf{A})$ to denote respectively the rth row and rth column of matrix **A**, and $r_r c_s$ to denote the result of multiplying corresponding elements of r_r and c_s and adding the results (*scalar product*), we can write (9) in the simpler form

$$\mathbf{AB} = \begin{bmatrix} r_1(\mathbf{A})c_1(\mathbf{B}) & r_1(\mathbf{A})c_2(\mathbf{B}) & \dots & r_1(\mathbf{A})c_n(\mathbf{B}) \\ r_2(\mathbf{A})c_1(\mathbf{B}) & r_2(\mathbf{A})c_2(\mathbf{B}) & \dots & r_2(\mathbf{A})c_n(\mathbf{B}) \\ \dots & \dots & \dots & \dots \\ r_n(\mathbf{A})c_1(\mathbf{B}) & r_n(\mathbf{A})c_2(\mathbf{B}) & \dots & r_n(\mathbf{A})c_n(\mathbf{B}) \end{bmatrix}. \tag{10}$$

From (10) it is obvious that in general $\mathbf{AB} \neq \mathbf{BA}$, so that the order of multiplication is important. In (10) we refer to **AB** as **B** *premultiplied* by **A**, or multiplied on the left by **A**, or as **A** *postmultiplied* by **B**, or multiplied on the right by **B**.

The transposed matrix of **A**, called $\mathbf{A}'$ or $\mathbf{A}^T$, is derived from **A** by interchanging rows and columns. If a matrix is *symmetric*, so that $a_{rs} = a_{sr}$, then $\mathbf{A}' = \mathbf{A}$, and

$$\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}'. \tag{11}$$

Other important cases in which the order of multiplication is immaterial are contained in the equations

$$\mathbf{AI} = \mathbf{IA}, \tag{12}$$

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{13}$$

The transpose of a product is given by

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}', \tag{14}$$

and its inverse by

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}; \tag{15}$$

note that in both operations the order of multiplication is reversed.

4. Associated with a matrix **A** is its *determinant*, denoted by det **A** or $|\mathbf{A}|$. Whereas the matrix is an array of numbers and can be regarded in many ways as an operator, the determinant is a pure number. For example,

$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}$$
$$= a_1(b_2c_3 - b_3c_2) - a_2(b_1c_3 - b_3c_1) + a_3(b_1c_2 - b_2c_1). \tag{16}$$

The sign associated with a_r is $(-)^{r+1}$, and the general rule for evaluation is obvious. The determinant

$$\begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix},$$

obtained by omitting the row and column containing a_1, is called a *minor* of order 2 of the original determinant.

It can be shown that the inverse $\mathbf{A}^{-1}$ of **A** is given by

$$\frac{1}{|\mathbf{A}|} \begin{bmatrix} A_{11} & -A_{21} & A_{31} & -A_{41} & \dots \\ -A_{12} & A_{22} & -A_{32} & A_{42} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}, \tag{17}$$

where the minors A_{rs} which are obtained by omitting from the original determinant the row and column containing a_{rs}, occur with alternate signs and are transposed in comparison with the corresponding elements a_{rs} of **A**.

5. When $|\mathbf{A}| = 0$, it is clear from (17) that the matrix **A** has no inverse. Such a matrix is called *singular*, and the corresponding linear equations have in general no solution. If $|\mathbf{A}| = 0$, the rows of **A** are not *linearly independent*; at least one can be obtained by linear combination of the others. For example, in the equations

$$\left.\begin{aligned} x_1+x_2+\ x_3 &= b_1, \\ x_1-x_2+2x_3 &= b_2, \\ 3x_1+x_2+4x_3 &= b_3, \end{aligned}\right\} \tag{18}$$

it can be verified that the determinant

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 3 & 1 & 4 \end{vmatrix} = 0,$$

and the third of (18) is obtained by adding twice the first to the second. In this case the equations are incompatible unless $2b_1+b_2 = b_3$, and if this holds we have effectively only two equations in three unknowns and there is an infinity of solutions.

If the equations are *homogeneous*, that is, the constants b_r are all zero, the equations have no solution other than $x_1 = x_2 = \ldots = x_n = 0$ unless the determinant vanishes, in which case we can omit one equation and solve the rest to find the ratios of the x_r, provided that the $n-1$ equations are not themselves linearly dependent. For example, in the homogeneous set of equations corresponding to (18), we can omit the last equation and solve

$$\left.\begin{aligned} x_1/x_3+x_2/x_3+1 &= 0, \\ x_1/x_3-x_2/x_3+2 &= 0, \end{aligned}\right\} \tag{19}$$

finding $x_1/x_3 = -1{\cdot}5$, $x_2/x_3 = 0{\cdot}5$, which also satisfy the remaining equation $3x_1/x_3+x_2/x_3+4 = 0$.

6. In solving a set of equations, it is a great advantage to be able to use the same number of decimal places throughout any one stage of the computation. It is a further advantage if the same number can be used at every stage. These advantages can be gained very simply by multiplying the various rows and columns of coefficients by powers of ten so as to make the largest coefficient in each row and column, including the column of constants, lie between 0·1 and 1·0. Multiplying the rows does not affect the solution, and multiplying the columns involves only a trivial change in the unknowns.

The same procedure should be carried out on a matrix of which the inverse is required. In this case, however, the multiplier of each row must subsequently be multiplied into the corresponding column of the inverse obtained, and the multiplier of each column into the corresponding row of the inverse, in order to recover the inverse of the original matrix.

With a symmetric matrix, when a method of solution or inversion is to be used which takes advantage of the symmetry, the multiplier of each row of the matrix must be the same as that of the corresponding column of the matrix, in order to maintain the symmetry. In general, it will then be possible to ensure only that the largest coefficients or elements lie in the wider range 0·1 to 10.

A matrix, modified in this way, is *well-conditioned* when its inverse, also, has its largest elements of order unity. In some cases, however, the elements of the inverse may have several figures before the decimal point, and it is then more difficult to get an accurate inverse or solution to the associated equations. Such a matrix or set of equations is said to be *ill-conditioned*, and this situation, which may be regarded as an approach towards singularity, manifests itself by a loss of significant figures during the computation.

SOLUTION OF EQUATIONS BY ELIMINATION OR PIVOTAL CONDENSATION

7. The simple elimination method taught at school is in practice carried out systematically and with the inclusion of frequent checks. If in equations (1) we select the largest of the coefficients of x_1, say a_{k1}, and add suitable multiples of the corresponding equation to all the other equations, so that in each resulting equation the coefficient of x_1 is zero, we shall be left with $n-1$ equations in the $n-1$ unknowns $x_2, x_3, \ldots, x_n$. The multipliers are clearly $-a_{11}/a_{k1}, -a_{21}/a_{k1}, \ldots$, and never exceed unity. The equation containing a_{k1}, the *pivot*, is called the *pivotal equation*, and is of course kept unchanged and temporarily left aside. We now select as pivot the largest coefficient of x_2 in the new set of $n-1$ equations and repeat the process. Continuing in this way we have finally a single equation in the unknown x_n. The various pivotal equations are then assembled, and have the form

$$\left.\begin{array}{rr} c_{11}x_1 + c_{12}x_2 + \ldots \qquad\qquad & + c_{1n}x_n = d_1, \\ c_{22}x_2 + c_{23}x_3 + \ldots \qquad & + c_{2n}x_n = d_2, \\ \multicolumn{2}{r}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ & c_{n-1,n-1}x_{n-1} + c_{n-1,n}x_n = d_{n-1}, \\ & c_{n,n}x_n = d_n. \end{array}\right\} \qquad (20)$$

The process used to produce this set of equations is known as *Gaussian elimination* or *pivotal condensation*.

We can now calculate x_n directly from the last of (20), and inserting its calculated value in the previous equation we can obtain x_{n-1}, and so on. This process is called *back-substitution*.

Several sets of equations, for the same **A** and varying **b**, can be solved almost simultaneously, as far as the elimination goes, by keeping several columns **b**; in particular, these may be the columns of the unit matrix if $\mathbf{A}^{-1}$ is required. Each back-substitution process is of course performed separately. In practice, the full equations at each stage are not recorded, but only the matrix of coefficients and column of constants.

The basic check on the elimination consists in carrying an extra column, whose rth element is formed of the sum of all the elements of $\mathbf{A}$ and $\mathbf{b}$ in the rth row. These elements are treated in the elimination exactly like the columns of constants, and after each elimination the 'sum' element should be equal, apart from small end-figure discrepancies through accumulation of rounding errors, to the sum of the other elements in its row.

The final results are checked by direct insertion in the original equations or, usually sufficiently, into an equation given by the sum of the original equations; this corresponds to the sum check in the elimination.

It is important to choose as pivot the largest element in a column; the multipliers are all then less than unity, and we can work with a constant number of decimals; see also Chapter 5.

If the matrix of coefficients is symmetric, symmetry is maintained if pivots are chosen on the diagonal. The work of elimination is then almost halved, but the multipliers may exceed unity.

EXAMPLE

8. The solution of the equations

$$\left.\begin{aligned} 0{\cdot}4096x_1+0{\cdot}1234x_2+0{\cdot}3678x_3+0{\cdot}2943x_4 &= 0{\cdot}4043,\\ 0{\cdot}2246x_1+0{\cdot}3872x_2+0{\cdot}4015x_3+0{\cdot}1129x_4 &= 0{\cdot}1550,\\ 0{\cdot}3645x_1+0{\cdot}1920x_2+0{\cdot}3728x_3+0{\cdot}0643x_4 &= 0{\cdot}4240,\\ 0{\cdot}1784x_1+0{\cdot}4002x_2+0{\cdot}2786x_3+0{\cdot}3927x_4 &= -0{\cdot}2557, \end{aligned}\right\}$$

is carried out as follows. The pivots are in italics, the sum column is labelled Σ, and the multipliers are called m. An extra decimal place is retained in the computations in order to compensate for the accumulation of rounding errors; compare Chapter 5, § 14.

Elimination

m	x_1	x_2	x_3	x_4	$\mathbf{b}$	Σ
	0·4096	0·1234	0·3678	0·2943	0·4043	1·5994
−0·54834	0·2246	0·3872	0·4015	0·1129	0·1550	1·2812
−0·88989	0·3645	0·1920	0·3728	0·0643	0·4240	1·4176
−0·43555	0·1784	0·4002	0·2786	0·3927	−0·2557	0·9942
−0·92230		0·31953	0·19982	−0·04848	−0·06669	0·40418(9)
−0·23723		0·08219	0·04550	−0·19759	0·06422	−0·00568(9)
		0·34645	0·11840	0·26452	−0·43179	0·29758√
			0·09062	−0·29245	0·33155	0·12972√
−0·19212			0·01741	−0·26034	0·16665	−0·07628(7)
				−0·20415	0·10295	−0·10120√

Back-substitution

x_1	x_2	x_3	x_4
−0·00593	−1·55547	2·03123	−0·50429

Check, using sum of original equations

$$1{\cdot}1771x_1+1{\cdot}1028x_2+1{\cdot}4207x_3+0{\cdot}8642x_4 = 0{\cdot}7276\ (0{\cdot}72761).$$

The loss of a significant figure in forming the third pivot indicates that the equations are somewhat ill-conditioned. The number of significant figures in this pivot is the maximum number that can be expected to be correct in the various elements of the solution, even though the last check may be better than this. The accurate solution of the equations to six decimal places is in fact

$$x_1 = -0{\cdot}006124, \quad x_2 = -1{\cdot}555598, \quad x_3 = 2{\cdot}031468, \quad x_4 = -0{\cdot}504263.$$

If the coefficients and constant terms are uncertain to the extent of half a unit in the last figure the solutions have even greater tolerances. A full analysis of the rounding errors in this example is given in Chapter 5, §§ 6–11.

If the pivots are selected at each stage from the largest coefficient in the complete relevant matrix, rather than from the columns in order, the tendency is for the pivots to lose significant figures gradually, and the last pivot is usually the smallest. This choice does not, however, lead to significantly greater accuracy in the final results.

Several variations of this straightforward elimination process are described in detail in [**16**].*

COMPACT ELIMINATION METHODS

9. For desk machines, the disadvantage of the simple elimination method is the large amount of recording; there is also an associated loss of accuracy, since at each recording a number is rounded and a small error introduced. This is avoided in the 'compact' elimination methods, of which we describe the method of Doolittle applied to the set of four equations

$$\left.\begin{aligned}
&\text{(i)} & a_{11}x_1+a_{12}x_2+a_{13}x_3+a_{14}x_4 &= b_1,\\
&\text{(ii)} & a_{21}x_1+a_{22}x_2+a_{23}x_3+a_{24}x_4 &= b_2,\\
&\text{(iii)} & a_{31}x_1+a_{32}x_2+a_{33}x_3+a_{34}x_4 &= b_3,\\
&\text{(iv)} & a_{41}x_1+a_{42}x_2+a_{43}x_3+a_{44}x_4 &= b_4.
\end{aligned}\right\} \qquad (21)$$

The procedure is as follows:

(*a*) Add a multiple of (i) to (ii) to eliminate x_1 from (ii), thus forming a new equation (ii).

(*b*) Add multiples of (i) and the new (ii) to (iii) to eliminate x_1 and x_2 from (iii), thus forming a new equation (iii).

(*c*) Add multiples of (i), the new (ii) and the new (iii) to (iv) to eliminate x_1, x_2 and x_3 from (iv), thus forming a new (iv).

The resulting equations (i), the new (ii), the new (iii) and the new (iv) have the form of (20), and can be solved as before by back-substitution. As before, a sum column is used as a check.

* Numbers in square brackets refer to the Bibliography on pages 145 to 165.

The computing sheet has the following appearance:

$$\begin{array}{ccccccccccc}
m_{12} & m_{13} & m_{14} & \quad & a_{11} & a_{12} & a_{13} & a_{14} & b_1 & \Sigma_1 \\
 & m_{23} & m_{24} & & & \alpha_{22} & \alpha_{23} & \alpha_{24} & \beta_2 & \Sigma_2 \\
 & & m_{34} & & & & \alpha_{33} & \alpha_{34} & \beta_3 & \Sigma_3 \\
 & & & & & & & \alpha_{44} & \beta_4 & \Sigma_4
\end{array}$$

The first multiplier m_{12} is obtained from the equation

$$m_{12}a_{11}+a_{21} = 0,$$

and the coefficients of the new (ii) are obtained from equations typified by

$$\alpha_{24} = m_{12}a_{14}+a_{24}.$$

The second column of multipliers is obtained from the equations

$$\begin{aligned} m_{13}a_{11}+a_{31} &= 0, \\ m_{13}a_{12}+m_{23}\alpha_{22}+a_{32} &= 0, \end{aligned}$$

and the coefficients of the new (iii) from equations typified by

$$\beta_3 = m_{13}b_1+m_{23}\beta_2+b_3.$$

Finally, the last column of multipliers is obtained from the equations

$$\begin{aligned} m_{14}a_{11}+a_{41} &= 0, \\ m_{14}a_{12}+m_{24}\alpha_{22}+a_{42} &= 0, \\ m_{14}a_{13}+m_{24}\alpha_{23}+m_{34}\alpha_{33}+a_{43} &= 0, \end{aligned}$$

and the coefficients of the new (iv) from equations typified by

$$\alpha_{44} = m_{14}a_{14}+m_{24}\alpha_{24}+m_{34}\alpha_{34}+a_{44}.$$

The final equations from which back-substitution is performed are not the same as the pivotal equations of the previous method, unless the chosen pivot at each stage in the latter is the first element in its column. In this event the two sets of equations are the same, except for rounding errors.

The saving in recording time and space is clear. The arrangement is also satisfactory in that quantities to be multiplied together lie in the same row of the computing sheet. The method of Crout, described in [**16**], has a still more compact lay-out but is less proof against error, since numbers which are to be multiplied together do not have this favourable combination of position.

In the case of symmetric matrices labour may be saved by computing the m's from the α's by means of the relation $m_{ij} = -\alpha_{ij}/\alpha_{ii}$.

METHODS DEPENDING ON MATRIX PROPERTIES

10. The matrix of coefficients c_{rs} in (20) is denoted by **U** and called *upper triangular*, since all its elements below the main diagonal are zero. A matrix with zero elements above this diagonal is labelled **L** and called *lower triangular*. Triangular matrices are obviously more convenient than complete matrices for solving linear equations: the determinant of such a matrix, moreover, is just the product of the diagonal terms.

With the Gaussian elimination method, the original equations (1), for which the matrix is complete, were transformed into equations (20), for which the matrix is upper triangular. The elimination is effectively equivalent to multiplying the original matrix **A** by a lower triangle **L**, producing an upper triangle **U**; thus

$$\mathbf{LA} = \mathbf{U}. \tag{22}$$

The equations from which the solutions are obtained by back-substitution are then

$$\mathbf{LAx} = \mathbf{Ux} = \mathbf{Lb}. \tag{23}$$

The matrix **L** here has ones in its diagonal; hence from (22) and the fact that the determinant of a matrix product is the product of the separate determinants [13], we see that the determinant of **A** is the same as that of **U**, and equal to the product of the diagonal terms of **U**. (If the pivots do not all lie on the diagonal, the sign of the determinant may be changed.)

Another class of method, the best for desk machines, uses the fact that a square matrix can be expressed as the product of two triangles, in the form

$$\mathbf{A} = \mathbf{LU}, \tag{24}$$

provided that it has non-zero leading principal minors, that is, the determinant composed of elements common to the first r rows and first r columns of **A** is non-zero for every $r = 1, 2, \ldots, n-1$. The diagonal terms of either **L** or **U** can here be chosen arbitrarily, the rest then being determined uniquely. If the matrix is symmetric, the diagonal of **U** is best taken to be the same as that of **L**; then **U** is the transpose of **L**, so that only one triangle has to be determined from the equation

$$\mathbf{A} = \mathbf{LL}', \tag{25}$$

though some elements will be imaginary if **A** is not positive definite (Chapter 3, § 3).

11. When multiplying two matrices on desk machines, it is best to record the transpose of the right-hand matrix vertically beneath the left-hand matrix, so that the rule (10) for multiplication can be written as

$$\mathbf{AB} = \begin{bmatrix} r_1(\mathbf{A})r_1(\mathbf{B}') & r_1(\mathbf{A})r_2(\mathbf{B}') & \cdots \\ r_2(\mathbf{A})r_1(\mathbf{B}') & r_2(\mathbf{A})r_2(\mathbf{B}') & \cdots \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \end{bmatrix},$$

and elements of rows are multiplied together, corresponding elements lying in the same column. In particular, if **B** is **A**$'$ we have

$$\mathbf{AA}' = \begin{bmatrix} \{r_1(\mathbf{A})\}^2 & r_1(\mathbf{A})r_2(\mathbf{A}) & \cdots \\ r_1(\mathbf{A})r_2(\mathbf{A}) & \{r_2(\mathbf{A})\}^2 & \cdots \\ \multicolumn{3}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \end{bmatrix}. \tag{26}$$

The determination of **L** and **U** in general is sufficiently illustrated by consideration of a matrix of order three. The notation and arrangement are as follows, the triangle **L** being here taken with unit diagonal elements, and the transpose of the upper triangle **U** being the lower triangle **U′**.

$$\begin{array}{cccccc}
 & & \mathbf{A} & & \mathbf{b} & \mathbf{\Sigma} \\
 & a_{11} & a_{12} & a_{13} & b_1 & \Sigma_1 \\
 & a_{21} & a_{22} & a_{23} & b_2 & \Sigma_2 \\
 & a_{31} & a_{32} & a_{33} & b_3 & \Sigma_3 \\
 & & \mathbf{L} & & & \\
 & 1 & & & & \\
 & l_{21} & 1 & & & \\
 & l_{31} & l_{32} & 1 & & \\
 & & \mathbf{U}' & & \mathbf{s} & \mathbf{x} \\
 & u_{11} & & & s_1 & x_1 \\
 & u_{12} & u_{22} & & s_2 & x_2 \\
 & u_{13} & u_{23} & u_{33} & s_3 & x_3 \\
\mathbf{y}' & y_1 & y_2 & y_3 & & \\
\mathbf{S} & S_1 & S_2 & S_3 & &
\end{array}$$

The method and order of calculation are as follows. The multiplication rule gives

$a_{11} = r_1(\mathbf{L})r_1(\mathbf{U}') = u_{11}$, $a_{12} = r_1(\mathbf{L})r_2(\mathbf{U}') = u_{12}$, $a_{13} = r_1(\mathbf{L})r_3(\mathbf{U}') = u_{13}$, giving the first column of **U′**;

$a_{21} = r_2(\mathbf{L})r_1(\mathbf{U}') = l_{21}u_{11}$, giving the second row of **L**;

$a_{22} = r_2(\mathbf{L})r_2(\mathbf{U}') = l_{21}u_{12}+u_{22}$, $a_{23} = r_2(\mathbf{L})r_3(\mathbf{U}') = l_{21}u_{13}+u_{23}$, giving the second column of **U′**;

$a_{31} = r_3(\mathbf{L})r_1(\mathbf{U}') = l_{31}u_{11}$, $a_{32} = r_3(\mathbf{L})r_2(\mathbf{U}') = l_{31}u_{12}+l_{32}u_{22}$, giving the third row of **L**; and finally

$a_{33} = r_3(\mathbf{L})r_3(\mathbf{U}') = l_{31}u_{13}+l_{32}u_{23}+u_{33}$, giving the last element in **U′**.

In the symmetric case **U′** is **L** and need not be recorded, the diagonal terms of **L** are denoted by l_{11}, l_{22} and l_{33}, and we have the equations

$$\begin{aligned}
a_{11} &= l_{11}^2, & a_{12} &= l_{11}l_{21}, & a_{13} &= l_{11}l_{31}, \\
 & & a_{22} &= l_{21}^2+l_{22}^2, & a_{23} &= l_{21}l_{31}+l_{22}l_{32}, \\
 & & & & a_{33} &= l_{31}^2+l_{32}^2+l_{33}^2,
\end{aligned}$$

for the successive determination of the l_{rs}.

12. When this *triangular resolution* or *decomposition* is finished, we can solve the linear equations (3) by two processes of back-substitution. Introducing the auxiliary vector **y**, defined by

$$\mathbf{Ux} = \mathbf{y}, \tag{27}$$

we can write

$$\mathbf{Ax} = \mathbf{LUx} = \mathbf{Ly} = \mathbf{b}, \tag{28}$$

solving for **y** from the last of (28), and for **x** from (27).

The elements of $\mathbf{y}$ are obtained in the same way as those of $\mathbf{U}'$. If they are written in transposed form as a row vector with components y_1, y_2, y_3, shown in position in the arrangement of § 11, then the equation $\mathbf{Ly} = \mathbf{b}$ gives

$$y_1 = b_1, \quad l_{21}y_1 + y_2 = b_2, \quad l_{31}y_1 + l_{32}y_2 + y_3 = b_3,$$

from which the y_r are obtained in succession.

As a check on the work we form the sum column $\mathbf{\Sigma}$, composed of the row sums of $\mathbf{A}$ and $\mathbf{b}$, and a sum row $\mathbf{S}$, composed of the column sums of $\mathbf{U}'$ and $\mathbf{y}'$. As each of the latter becomes available we use the successive relations

$$r_1(\mathbf{L})\,\mathbf{S} = \Sigma_1, \quad r_2(\mathbf{L})\,\mathbf{S} = \Sigma_2, \quad r_3(\mathbf{L})\,\mathbf{S} = \Sigma_3,$$

or

$$S_1 = \Sigma_1, \quad l_{21}S_1 + S_2 = \Sigma_2, \quad l_{31}S_1 + l_{32}S_2 + S_3 = \Sigma_3.$$

We finally calculate $\mathbf{x}$ from (27) from equations typified by

$$c_r(\mathbf{U}')\,\mathbf{x} = y_r. \tag{29}$$

The elements of $\mathbf{x}$ are recorded as shown, and calculated from the successive equations obtained by taking $r = 3, 2, 1$ in (29), and given by $u_{33}x_3 = y_3$, $u_{23}x_3 + u_{22}x_2 = y_2$, $u_{13}x_3 + u_{12}x_2 + u_{11}x_1 = y_1$. If $\mathbf{s}$ is the column formed of the row sums of $\mathbf{U}'$, a suitable check on this back-substitution is given by

$$s_1x_1 + s_2x_2 + s_3x_3 = y_1 + y_2 + y_3.$$

It should be noticed that the array of multipliers m_{ij} in the Doolittle method (§ 9) is the same as $\mathbf{L}'$, with the signs changed and the unit diagonal terms omitted, and that the array of coefficients a_{ij} and α_{ij} in Doolittle's resulting equations is the same as $\mathbf{U}$. Thus the computations are precisely the same in the two methods; only the arrangement differs.

If the matrix is symmetric, equations (27) and (28) are replaced by

$$\left.\begin{aligned} \mathbf{L}'\mathbf{x} &= \mathbf{y}, \\ \mathbf{Ax} = \mathbf{LL}'\mathbf{x} &= \mathbf{Ly} = \mathbf{b}, \end{aligned}\right\} \tag{30}$$

so that $\mathbf{U}' = \mathbf{L}$ and the only change in arrangement is the complete omission of $\mathbf{U}'$, the sums $\mathbf{S}$ and $\mathbf{s}$ being attached to $\mathbf{L}$.

13. The solution by this method of the previous example is given on the next page. The fact that the answers agree with those of § 8 to barely four decimals is due to the ill-conditioning of the equations noted previously.

14. For matrix inversion there are various possibilities following the triangular resolution. One method is to invert both $\mathbf{L}$ and $\mathbf{U}$, and then find $\mathbf{A}^{-1}$ from

$$\left.\begin{aligned} \mathbf{A}^{-1} &= \mathbf{U}^{-1}\mathbf{L}^{-1} \text{ (unsymmetric case),} \\ \mathbf{A}^{-1} &= (\mathbf{L}')^{-1}\mathbf{L}^{-1} \text{ (symmetric case).} \end{aligned}\right\} \tag{31}$$

In the second of (31), $\mathbf{L}^{-1} = \{(\mathbf{L}')^{-1}\}'$, so that only one triangle has to be inverted.

The arrangement for inversion of triangles and the final multiplication are described in [**16**]; still more compact methods are described in [**17**].

These compact methods, in which at most one triangle is inverted, are usually preferred.

A				b	Σ
0·4096	0·1234	0·3678	0·2943	0·4043	1·59940√
0·2246	0·3872	0·4015	0·1129	0·1550	1·28120(19)
0·3645	0·1920	0·3728	0·0643	0·4240	1·41760(59)
0·1784	0·4002	0·2786	0·3927	−0·2557	0·99420√

L			
1			
0·54834	1		
0·88989	0·25721	1	
0·43555	1·08426	16·65290	1

	U′				s	x
	0·40960				0·40960	−0·00609
	0·12340	0·31953			0·44293	−1·55560
	0·36780	0·19982	−0·00590		0·56172	2·03144
	0·29430	−0·04848	−0·18513	3·40003	3·46072	−0·50427
y′	0·40430	−0·06669	0·08137	−1·71453	(−1·29555√)	
S	1·59940	0·40418	−0·10966	1·68550		

2

LINEAR EQUATIONS AND MATRICES: DIRECT METHODS ON AUTOMATIC COMPUTERS

INTRODUCTION

1. When linear equations are solved on a desk machine, the minimization of the number of quantities which have to be written down and the convenience of the layout are of paramount importance. The time taken to write a number is comparable with that taken to perform an arithmetical operation, and a high percentage of the total number of mistakes occurs in the writing stage. Although properly applied checks give complete protection against undetected errors, the solution of a system of equations of quite moderate order is tedious unless mistakes are infrequent.

In contrast, an automatic computer in good working order may be relied upon to solve a very large system of equations without making any mistakes. The problems of layout are still present, though in a rather different form. They are now mainly concerned with the use of an auxiliary store having an access time different from that of the high-speed store. For instance, it is often of decisive importance whether matrices are held in the auxiliary store in rows or in columns.

2. In other respects, an automatic computer is less flexible than a desk machine. For example, the method of triangular decomposition described in §§ 10–14 of the previous chapter is superior to the elimination method only if the scalar products are accumulated without rounding each contribution. It is therefore essentially a fixed-point method, but an experienced hand computer will add extra figures when the need arises. In practice this need is likely to arise quite frequently, because no provision has been made for the selection of pivots comparable with that described in Gaussian elimination. In consequence, a diagonal element of the upper triangular matrix **U** may sometimes be quite small, with the result that some of the elements of **L** and **U** are considerably larger than those of the matrix of the original equations.

Unsystematic modifications of this type are unsatisfactory in automatic work. In general, in order to achieve the maximum accuracy in the solution we try to design programmes in such a way that all numbers fully occupy the storage registers (perhaps it should be added that most automatic computers do not work any faster with numbers which are less than a full word length). Accordingly, in this chapter we shall analyse the computing procedures in more detail.

3. The precise details of the arithmetical facilities of the computer are of great importance. On many machines both fixed-point and floating-point facilities are available. If floating-point operations are used throughout, then there is apparently no need to give detailed attention to the size of numbers arising during the course of the computation. It is shown in Chapter 5, however, that this procedure does not remove the necessity for interchanges in Gaussian elimination and related methods. Moreover, for a given word length the precision of floating-point arithmetic is lower than that attainable with fixed-point arithmetic because some digits have to be allocated to the exponent. In most matrix problems it is possible to carry out the necessary scaling in fixed-point operations using fewer digits than would be required for the floating-point exponents.

4. Fixed-point operation is particularly advantageous on computers which are equipped with the following two facilities.

(i) The ability to accumulate scalar products, $\Sigma a_i b_i$, exactly. This facility is possessed by all desk machines; the exact scalar product is usually produced whether required or not. On automatic computers the decisive feature is whether or not double precision in the product may be obtained without special programming. The main advantage of this facility is that only one rounding error is made, this occurring when the accumulated sum is rounded to single precision.

(ii) The ability to divide a double-precision number by a single-precision number. When this facility is provided in conjunction with the facility (i), only one rounding error is introduced in the computation of quantities of the form $(\Sigma a_i b_i)/c$. Both facilities are available on the computers in use at N.P.L., and their provision elsewhere is becoming increasingly common.

GAUSSIAN ELIMINATION

5. *Elimination with selection of pivots* may be carried out quite satisfactorily with fixed-point arithmetic. Storage presents no special problems because the successive reduced equations may be overwritten in the locations occupied by the original matrix. It is usually desirable to design the programme so that it will deal simultaneously with any number, r say, of right-hand sides. For a system of order n, a total of $n(n+r)$ storage locations is then required.

6. Since pivotal selection is used, all the multipliers are bounded by unity and may therefore be computed without scale factors. If the original equations are scaled so that all coefficients and constant terms are numerically less than $\frac{1}{8}$, say, it is then extremely uncommon for an element of any of the reduced equations to exceed unity.

In a general-purpose programme it is desirable to include some procedure which deals automatically with the danger of 'overspill'. The following method is based on the observation that if all the elements of a given reduced set of equations are less than $\frac{1}{2}$, then no element of the next reduced set can exceed unity. As each reduced equation is formed, its elements are tested for size. If any exceeds $\frac{1}{2}$, then that equation is divided throughout by 2. In this way the maximum precision is preserved and the possible need to repeat the computations as a result of overspill is avoided.

7. In the programmes written for the N.P.L. computers, the elements of the rth pivotal row are interchanged with those of the rth row at each stage of the reduction so that the final triangular set of equations is stored in the natural order. For this reason elimination with selection of pivots is often referred to as *elimination with interchanges*. On other machines it is sometimes more convenient to leave each row in its original position and to store the locations of the pivotal rows. The location of the pivotal equation for the $(r+1)$th reduction may be determined during the course of the rth reduction; we merely have to keep a record of the size and location of the numerically largest coefficient of x_{r+1} occurring in the reduced equations up to and including the current stage.

THE BACK-SUBSTITUTION

8. In the reduction to triangular form, the need for scaling is almost non-existent. In the back-substitution, however, scaling is the major preoccupation in fixed-point work, since the scaling of the original equations places no restriction whatever on the size of the solutions. A convenient scheme is the following.

The rth pivotal equation has the form

$$a^{(r)}_{r,r}x_r + a^{(r)}_{r,r+1}x_{r+1} + \dots + a^{(r)}_{r,n}x_n = b^{(r)}_r. \tag{1}$$

In applying this equation, $x_{r+1}, x_{r+2}, \dots, x_n$ have already been calculated; x_r is now found from

$$x_r = \frac{-a^{(r)}_{r\,r+1}x_{r+1} - a^{(r)}_{r,r+2}x_{r+2} - \dots - a^{(r)}_{r,n}x_n + b^{(r)}_r}{a^{(r)}_{r,r}}, \tag{2}$$

and if the facilities described in § 4 are available, the computation of x_r involves only one rounding error. The process may be carried out in this simple form until a calculated x_r exceeds unity. When this happens, a scale factor 2^{-s}, where $s \geqslant 1$, is introduced to bring x_r into the permissible range. All the previously calculated x_i are then multiplied by this factor. Scale factors may be introduced in this way several times during the back-substitution. At the stage when the accumulated scale factor is 2^{-s}, the equation determining x_r may be written in the form

$$x_r = \frac{-a^{(r)}_{r\,r+1}x_{r+1} - \dots - a^{(r)}_{r\,n}x_n - b^{(r)}_r(-2^{-s})}{a^{(r)}_{r,r}}, \tag{3}$$

where the x_i denote the currently stored values of these variables. If the quantity (-2^{-s}) is stored as though it were an extra variable, then the numerator of (3) may be treated as a scalar product of order $n+1-r$. At the end of the back-substitution the stored value of 2^{-s} gives the final scale factor. The computed solution $\mathbf{x}$ will therefore have its maximum component in the range $(2^{s-1}, 2^s)$ and with the full number of significant figures, unless no scaling has taken place, in which event all components of $\mathbf{x}$ are numerically less than unity.

9. In many computations we require the solution of a system of equations with a number of different right-hand sides, not all of which are known at the time when the elimination is performed. For example,

with some methods of computing latent vectors (see Chapter 3) we wish to solve successively the systems

$$\mathbf{A}\mathbf{x}^{(r+1)} = \mathbf{b}^{(r)}, \qquad \mathbf{b}^{(r)} = k^{(r)}\mathbf{x}^{(r)} \quad (r = 1, 2, \ldots), \tag{4}$$

where the $k^{(r)}$ are scalars, usually normalizing factors. Here each right-hand side is determined by the previous solution.

To cope with such problems, sufficient information must be stored so that we can apply to the right-hand side the transformations which are normally applied during the reduction process and the back-substitution. In order to be able to do this we must store the multipliers, the pivotal rows and the details of the interchanges.

10. We now describe a convenient way of doing this which requires a total storage space of $n(n+1)$ words. The configuration at a typical stage in the reduction is illustrated below for a matrix of order 5, at the end of the second stage. The quantities m_{i1} and m_{i2} are the multipliers used in the first and second stages, and the a_{ij} denote the current values of the coefficients of the equations. The multipliers have the opposite signs to those introduced in § 8 of Chapter 1.

$$\begin{array}{cccccc}
p_1 & a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\
p_2 & m_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\
p_3 & m_{31} & m_{32} & a_{33} & a_{34} & a_{35} \\
0 & m_{41} & m_{42} & a_{43} & a_{44} & a_{45} \\
0 & m_{51} & m_{52} & a_{53} & a_{54} & a_{55}
\end{array}$$

Each element is given the suffixes corresponding to the storage position it occupies; no account of interchanges is taken in this nomenclature. The quantities p_1 and p_2 are the numbers of the pivotal rows in the first and second stages, and are therefore integers less than 6. The first and second pivotal rows are now of course in the first and second positions. During the second reduction the number, p_3, of the next pivotal row will have been determined in advance as described in § 7.

The elements a_{ij} in row p_3 are now interchanged with those in row 3, but the m_{ij} in these rows must not be interchanged. Multiples m_{43} and m_{53} of the new row 3 are subtracted from the current rows 4 and 5, the new values of a_{ij} are overwritten on the elements from which they are derived, and m_{43} and m_{53} are overwritten on a_{43} and a_{53}, respectively. During this reduction the next pivotal-row number, p_4, is determined, and is written at the beginning of row 4. We are then ready to start the next stage of the reduction.

11. The use of the stored information to deal with a right-hand side $\mathbf{b}$ may be adequately described by the reduction to the third stage for the system of order 5. The right-hand side will occupy 5 storage locations, and we call the current contents of these locations, $b_1, b_2, \ldots, b_5$. The third stage is as follows.

(i) Interchange b_3 and b_{p_3}. (If $p_3 = 3$ no interchange is necessary, but it is simplest to allow the programme to effect the 'interchange'.)

(ii) Subtract multiples m_{43}, m_{53} of b_3 from b_4 and b_5 and overwrite the new values in the 4th and 5th locations. Note that the elements b_1, b_2 obtained in the earlier stages are unaltered since p_3 is not less than 3.

In general, there are $n-1$ stages of this type and when they are completed, the equations are ready for the back-substitution.

The complete processing of a right-hand side, starting from the stage at which the elimination has been completed, requires approximately n^2 multiplications and n divisions. If the process of division is slower than that of multiplication and there are several right-hand sides, it will be economical to form the reciprocals of the pivotal elements. If we were to compute the inverse of $\mathbf{A}$, then we could obtain the solution corresponding to a given right-hand side, $\mathbf{b}$, by forming the product $\mathbf{A}^{-1}\mathbf{b}$. This requires n^2 multiplications. It must be remembered, however, that the number of multiplications needed to calculate the inverse of $\mathbf{A}$ exceeds that needed to produce the matrix of multipliers and pivotal rows by about $\frac{2}{3}n^3$. Therefore unless we have a *very* large number of right-hand sides, it is uneconomical to find the inverse.

Programmes of the type which decompose the matrix of coefficients into two triangular matrices, which may then be used to solve with any right-hand side, have proved to be among the most useful of those in the N.P.L. library.

VARIANTS OF GAUSSIAN ELIMINATION

12. Many methods of solving linear equations have been devised which are essentially variants of Gaussian elimination. Often they exploit some special feature of a particular machine. Most commonly they are designed to compensate for the loss of speed when the system of equations is too large to be held in the high-speed store. For the most part they have no specific names but they are more important in automatic work than many of the named variants used on desk machines. Furthermore, they are often mathematically distinct from Gaussian elimination and triangular decomposition, whereas many of the named variants used on desk machines differ only in the layout of the work.

13. There is one variant which is of particular advantage when the system of equations is fed in as data and not produced by the computer itself, since it enables the solution to be effected using approximately $\frac{1}{2}n^2$ locations instead of n^2. There are n major steps in this process. At a typical step, the rth, there are $r-1$ equations in the high-speed store, the first involving $x_1, x_2, \ldots, x_n$, the second $x_2, x_3, \ldots, x_n$, and the $(r-1)$th, $x_{r-1}, x_r, \ldots, x_n$. At this stage only $r-1$ equations have been read into the store. For $r = 4$, $n = 6$ and one right-hand side, the stored system has the following configuration:

$$\begin{array}{ccccccc}
a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & b_1 \\
 & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & b_2 \\
 & & a_{33} & a_{34} & a_{35} & a_{36} & b_3
\end{array}$$

We refer to the reduced equations represented by the three rows as the 1st, 2nd and 3rd equations respectively. Interchanges may take place in the next step but we shall still refer to the equation currently occupying the 1st position as the 1st equation, and similarly for the others. The 4th step proceeds as follows.

(i) The 4th equation is read into the store.

(ii) The coefficients of x_1 in the 1st and 4th equations are compared and if necessary the equations are interchanged so that the 1st contains the larger coefficient. A multiple of the 1st equation is then subtracted from the 4th, to give a zero coefficient of x_1. This multiple cannot exceed unity.

(iii) The new 4th equation is now compared with the 2nd. If necessary, an interchange is performed so that the 2nd equation has the larger coefficient of x_2. A multiple of the 2nd equation is now subtracted from the 4th to give a zero coefficient of x_2.

(iv) The 4th equation is now compared with the 3rd and, if necessary, an interchange is performed so that the 3rd equation has the larger coefficient of x_3. A multiple of the 3rd equation is then subtracted from the 4th to give a zero coefficient of x_3. This completes the 4th stage.

It is evident that the total storage required for a system with one right-hand side is only slightly greater than that needed to hold the final triangular set of equations and is thus about half that needed to hold the full set of equations.

14. We cannot take advantage of this variant if we wish to solve with right-hand sides which are not known when the elimination is performed. In this case we will need to store the multipliers as well as the pivotal rows and will therefore require n^2 storage locations. However, the method of performing the interchanges still has advantages on some computers, particularly for sets of equations which are large enough to require the use of the auxiliary store. Furthermore, the details of the interchanges can now be stored in a particularly simple manner. We record whether or not an interchange took place just before each of the multipliers was produced, by sacrificing the least significant digit of each multiplier and storing in its place a one or a zero.

TRIANGULAR DECOMPOSITION WITH INTERCHANGES

15. Interchanges may be introduced in the method of triangular decomposition in such a way that all the elements of $\mathbf{L}$ remain less than unity. Indeed, this is just as important for triangular decomposition as for Gaussian elimination. A satisfactory technique has not been described in the literature, however, no doubt because of the inconvenience of carrying out the interchanges on a desk machine. We now describe a procedure which has been programmed for the ACE computer at N.P.L.

16. The matrix of the coefficients is processed column by column. There are n major steps, the rth of which is concerned with the modification of the rth column only. In addition to the storage space occupied by the matrix, n pairs of storage locations are required to hold n exact scalar products accumulated during the column processing. The configuration at the beginning of the rth step is typified by that shown below for $r = 3$, $n = 5$.

$$\begin{array}{cccccc}
u_{11} & u_{12} & a_{13} & a_{14} & a_{15} & s_1s_1 \\
l_{21} & u_{22} & a_{23} & a_{24} & a_{25} & s_2s_2 \\
l_{31} & l_{32} & a_{33} & a_{34} & a_{35} & s_3s_3 \\
l_{41} & l_{42} & a_{43} & a_{44} & a_{45} & s_4s_4 \\
l_{51} & l_{52} & a_{53} & a_{54} & a_{55} & s_5s_5 \\
p_1 & p_2 & 0 & 0 & 0 &
\end{array}$$

In the general case, the elements of the first $r-1$ columns of **L** and **U** (apart from the diagonal elements of **L** which are unity and are not stored) have been produced and overwritten on the corresponding elements of **A**. The quantity p_t appearing at the foot of the ith column is an integer specifying the interchanges and is defined below. The pairs of registers needed to store the double-length scalar products are denoted by $s_i s_i$. To avoid misunderstanding we stress that the triangular matrices **L** and **U**, stored in the machine at the conclusion of the process are *not* such that **LU** is equal to **A** with its rows rearranged. For this to be true the elements in each column of **L** would have to be permuted and a different permutation used for each column. Each column of the stored matrix **L** is in the most convenient form for the processing of the columns of **A** and of a right-hand side.

The rth step takes place in the following four stages:

(i) Multiplication of each element of the rth column by unity to give n double-precision numbers, and the storage of the results in the double locations ss.

(ii) Successive calculation of $u_{1,r}, u_{2,r}, \ldots, u_{r-1,r}$.

(iii) Calculation of p_r and $u_{r,r}$.

(iv) Calculation of $l_{r+1,r}, l_{r+2,r}, \ldots, l_{n,r}$.

Stage (i) is self-explanatory.

Stage (ii) has $r-1$ substages in the tth of which $u_{t,r}$ is computed. The tth step is as follows. Extract s_{p_t} and round to single precision to give $u_{t,r}$. Overwrite this on $a_{t,r}$ and then overwrite s_t on s_{p_t}. Subtract multiples $l_{t+1,t}, l_{t+2,t}, \ldots, l_{n,t}$ of $u_{t,r}$ from $s_{t+1}, s_{t+2}, \ldots, s_n$ and overwrite the modified scalar products so obtained on their old values.

Stage (iii). Select the largest, s_{p_r}, of the scalar products $s_r, s_{r+1}, \ldots, s_n$. This s_{p_r} defines p_r which is entered at the foot of the rth column. Round s_{p_r} to give $u_{r,r}$, and overwrite on $a_{r,r}$. Overwrite s_r on s_{p_r}.

Stage (iv) has $n-r$ substages in the tth of which $l_{r+t,r}$ is computed by dividing $s_{r+t,r}$ by $u_{r,r}$ to give $l_{r+t,r}$, and then overwriting the result on $a_{r+t,r}$.

It is evident that only one rounding error is made when calculating each element of **L** and **U**, provided that the computer has both of the facilities described in § 4.

Since we have stored complete information on the triangular decomposition, we may subsequently deal with any number of right-hand sides. The details of the processing of the right-hand side closely follow those of the decomposition, and we omit them.

17. It may be noted that the exact accumulation of scalar products is not essential to the columnwise processing; it merely improves the accuracy. The method may be used without this facility and is then *exactly equivalent* to performing Gaussian elimination with interchanges.

ILL-CONDITIONED EQUATIONS

18. For a system of ill-conditioned equations the calculated solution may not be sufficiently close to the exact solution owing to the limitation of the working accuracy by the given word-length. If we have stored complete information about the matrix transformation, an improved

solution may be obtained without recourse to double-length arithmetic as follows. Let $\mathbf{x}^{(1)}$ be the computed solution of the equations

$$\mathbf{Ax} = \mathbf{b}. \tag{5}$$

If we define $\mathbf{r}^{(1)}$ by the equation

$$\mathbf{r}^{(1)} = \mathbf{b} - \mathbf{Ax}^{(1)}, \tag{6}$$

then we have

$$\mathbf{A}[\mathbf{x} - \mathbf{x}^{(1)}] = \mathbf{r}^{(1)}. \tag{7}$$

The correction to be applied to $\mathbf{x}^{(1)}$ to give the solution of (5) is the solution of equation (7). Now unless the matrix $\mathbf{A}$ is very ill-conditioned, the components of $\mathbf{r}^{(1)}$ will be smaller than those of $\mathbf{b}$. On a computer which accumulates double-length scalar products we may calculate $\mathbf{r}^{(1)}$ exactly; if its largest element lies between $2^{-k^{(1)}}$ and $2^{-(k^{(1)}+1)}$ we may multiply all its components by $2^{k^{(1)}}$, and round the resulting vector to single precision. We denote this vector by $2^{k^{(1)}}\bar{\mathbf{r}}^{(1)}$ and solve the equations

$$\mathbf{A}\boldsymbol{\delta}^{(1)} = 2^{k^{(1)}}\bar{\mathbf{r}}^{(1)}, \tag{8}$$

using the stored information, to obtain $\boldsymbol{\delta}^{(1)}$. The vector

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + 2^{-k^{(1)}}\boldsymbol{\delta}^{(1)}$$

is then an improved solution. We may continue this process to obtain sequences $\mathbf{x}^{(i)}$, $\mathbf{r}^{(i)}$, $k^{(i)}$, $\bar{\mathbf{r}}^{(i)}$ and $\boldsymbol{\delta}^{(i)}$ defined by

$$\mathbf{r}^{(i)} = \mathbf{b} - \mathbf{Ax}^{(i)}; \tag{9}$$

the maximum element of $2^{k^{(i)}}\bar{\mathbf{r}}^{(i)}$ lies between $\frac{1}{2}$ and 1, and

$$\mathbf{A}\boldsymbol{\delta}^{(i)} = 2^{k^{(i)}}\bar{\mathbf{r}}^{(i)}, \tag{10}$$

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + 2^{-k^{(i)}}\boldsymbol{\delta}^{(i)}. \tag{11}$$

19. Three comments may be made on the above process.

(i) The exact accumulation of scalar products is important in the computation of $\mathbf{r}^{(i)}$. If each multiplication is rounded individually or if the residuals are computed using single-precision floating-point arithmetic, the error made in computing the residual may be comparable with its true value. If floating-point arithmetic is used we must work to double precision to obtain residuals of comparable accuracy to those obtained on a fixed-point computer equipped with the facility of accumulation.

(ii) If the equations are too ill-conditioned, then although the components of $\mathbf{r}^{(1)}$ may be much smaller than those of $\mathbf{b}$, the $\mathbf{r}^{(i)}$ will not decrease with each iteration. They will either remain much the same size, or even increase. In either case we will not gain accuracy by repeating the process and it is almost certain that $\mathbf{x}^{(1)}$ has no correct figures. In this event it is necessary to repeat the whole process using double-length arithmetic throughout.

(iii) If only one or two binary figures are gained per stage, then many iterations are needed before the answers are correct to single precision. If the solution is required for a number of different right-hand sides, then it will be more economical to perform the computation in double-precision arithmetic. If single-precision computation with a t-digit word is capable of giving a solution at all, double-precision computation will give at least t correct figures in the first step. If single-precision work provides no solution, then double-precision work will be required anyway. If there is reason to expect the equations to be very ill-conditioned, the case for undertaking double-precision work at the outset is very strong.

The main advantage of using the iterative process to improve the solutions is that it provides an extremely reliable indication of the accuracy of the solution. If $\mathbf{x}^{(1)}$ agrees with $\mathbf{x}^{(2)}$ to r figures and $\mathbf{x}^{(2)}$ with $\mathbf{x}^{(3)}$ to $2r$ figures, we may be fairly certain that $\mathbf{x}^{(3)}$ has $3r$ correct figures.

It should be emphasized that the extra figures obtained in this way will be meaningful only if the original equations are exact or if the errors in them are correlated in some special way. This is discussed further in Chapter 5.

3

LATENT ROOTS AND VECTORS OF MATRICES

INTRODUCTION

1. The problem of finding the latent roots of a matrix is of fundamental importance and frequent occurrence in numerical analysis; it may be defined as follows.

Given a square matrix $\mathbf{A}$ we require those values of λ for which the set of linear equations

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{1}$$

has a non-trivial solution $\mathbf{x}$. The values of λ are called the *latent roots* or *eigenvalues* of $\mathbf{A}$ and the corresponding vectors $\mathbf{x}$ are called the *latent vectors* or *eigenvectors* of $\mathbf{A}$. Each vector $\mathbf{x}$ is determined apart from an arbitrary constant multiplier. It is usual to choose the multiplier so that the sum of the squares of the components of $\mathbf{x}$ is unity, or sometimes so that its largest element is unity. Such a vector is called a *normalized* latent vector.

Quite commonly the latent root problem arises in the form

$$\mathbf{C}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}, \tag{2}$$

but this may be converted to the simpler form by writing it as

$$\mathbf{B}^{-1}\mathbf{C}\mathbf{x} = \lambda\mathbf{x}.$$

2. The latent root problem often arises from the solution of simultaneous linear differential equations with constant coefficients. If, for example, we have a set of n second-order equations we may write these in vector form as

$$\mathbf{A}\ddot{\mathbf{x}} + \mathbf{B}\dot{\mathbf{x}} + \mathbf{C}\mathbf{x} = 0, \tag{3}$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are $n \times n$ matrices and the dot represents differentiation. Introducing n new variables $\mathbf{p}$ defined by

$$\dot{\mathbf{x}} = \mathbf{p}, \tag{4}$$

we may write the equations in the form

$$\mathbf{A}\dot{\mathbf{p}} + \mathbf{B}\mathbf{p} + \mathbf{C}\mathbf{x} = 0. \tag{5}$$

From the theory of differential equations it is known that the solution of the set of equations (4) and (5) in $2n$ unknowns, $\mathbf{x}$ and $\mathbf{p}$, consists of linear combinations of solutions of the type

$$\mathbf{x} = \mathbf{a}e^{\lambda t}, \quad \mathbf{p} = \mathbf{b}e^{\lambda t},$$

where $$\lambda \mathbf{a} = \mathbf{b}, \tag{6}$$

in virtue of (4), and (5) gives the relation

$$\mathbf{A}\lambda\mathbf{b}+\mathbf{B}\mathbf{b}+\mathbf{C}\mathbf{a} = 0. \tag{7}$$

Equations (6) and (7) may be combined into the single matrix equation

$$\lambda \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbf{I} \\ -\mathbf{C} & -\mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \tag{8}$$

where $\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}$ is a vector with $2n$ components. Renaming this vector $\mathbf{z}$, we may write (8) in the form

$$\lambda \mathbf{P}\mathbf{z} = \mathbf{Q}\mathbf{z}, \tag{9}$$

equivalent to that of equation (2).

3. An important physical problem which gives rise to a latent root problem is the determination of the periods of free vibration of a dynamical system about a position of equilibrium, for which the differential equations are of the form

$$\mathbf{A}\ddot{\mathbf{x}} = -\mathbf{B}\mathbf{x}. \tag{10}$$

Here the matrices **A** and **B** are both *positive definite*, that is, the *quadratic form*

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \Sigma a_{ij} x_i x_j \tag{11}$$

is positive for all real values of the variables x_i; similarly for **B**. A necessary and sufficient condition for a *symmetric* matrix to be positive definite is that all of its latent roots are positive [**14**]. For positive definite **A** and **B** it can be shown that the solutions of equation (10) are of the form $\mathbf{x} = \mathbf{y}\, e^{i\lambda t}$, where the λ are real and satisfy the equation

$$\mathbf{A}\lambda^2\mathbf{y} = \mathbf{B}\mathbf{y}.$$

FUNDAMENTAL RELATIONS

4. The latent roots of a matrix **A** are those values of λ for which

$$|\mathbf{A}-\lambda\mathbf{I}| = 0. \tag{12}$$

From (12) we could obtain an explicit polynomial equation of degree n for λ which is called the *characteristic equation* of **A**. We may write this as

$$a_0+a_1\lambda+a_2\lambda^2+\ldots+a_{n-1}\lambda^{n-1}+(-1)^n\lambda^n = 0. \tag{13}$$

Equation (13) has n roots $\lambda_1, \lambda_2, \ldots, \lambda_n$, and these are the latent roots of the matrix **A**. They may be real or complex; for simplicity we assume that they are distinct. The corresponding latent vectors will be called $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$; they are linearly independent if the latent roots are distinct [**14**].

The matrix **A**$'$, which is the transposed matrix of **A**, will have as its latent roots those values of λ for which

$$|\mathbf{A}'-\lambda\mathbf{I}| = 0. \tag{14}$$

Since $(\mathbf{A}-\lambda\mathbf{I})' = \mathbf{A}'-\lambda\mathbf{I}$ and the determinant of a matrix is equal to the determinant of its transpose, $\mathbf{A}'$ has the same latent roots as $\mathbf{A}$, though in general its latent vectors will be different. The latent vectors of $\mathbf{A}'$ corresponding to $\lambda_1, \ldots, \lambda_n$ will be denoted by $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$.

We then have
$$\mathbf{A}\mathbf{x}_r = \lambda_r \mathbf{x}_r \tag{15}$$
and
$$\mathbf{A}'\mathbf{x}_s^* = \lambda_s \mathbf{x}_s^*. \tag{16}$$
The transpose of (15) gives the equation
$$\mathbf{x}_r'\mathbf{A}' = \lambda_r \mathbf{x}_r', \tag{17}$$
premultiplication of (16) by $\mathbf{x}_r'$ gives
$$\mathbf{x}_r'\mathbf{A}'\mathbf{x}_s^* = \lambda_s \mathbf{x}_r'\mathbf{x}_s^*, \tag{18}$$
and postmultiplication of (17) by $\mathbf{x}_s^*$ gives
$$\mathbf{x}_r'\mathbf{A}'\mathbf{x}_s^* = \lambda_r \mathbf{x}_r'\mathbf{x}_s^*. \tag{19}$$
From (18) and (19) we find
$$0 = (\lambda_r - \lambda_s)\mathbf{x}_r'\mathbf{x}_s^*, \tag{20}$$
so that
$$\mathbf{x}_r'\mathbf{x}_s^* = 0 \quad \text{if} \quad \lambda_r \neq \lambda_s. \tag{21}$$
The two sets of vectors $\mathbf{x}_i$ and $\mathbf{x}_i^*$ are said to be *biorthogonal.* If $\mathbf{A}$ is symmetric then $\mathbf{A} = \mathbf{A}'$, the $\mathbf{x}_i$ and $\mathbf{x}_i^*$ coincide, and we have
$$\mathbf{x}_i'\mathbf{x}_j = 0 \quad (i \neq j). \tag{22}$$

The latent roots of a symmetric matrix are real. For if λ is a complex latent root, $\bar{\lambda}$ is also a latent root and is different from λ. If $\mathbf{x}, \bar{\mathbf{x}}$ denote the corresponding latent vectors, equation (22) shows that the scalar product of $\mathbf{x}$ and $\bar{\mathbf{x}}$ is zero. This is clearly impossible since this scalar product is the sum of the squares of the real and imaginary parts of the elements of $\mathbf{x}$.

ITERATIVE PROCESSES

5. One of the simplest methods of finding simultaneously a latent root and vector of a matrix $\mathbf{A}$ is the following. Suppose the latent roots $\lambda_1, \lambda_2, \ldots, \lambda_n$, assumed real and distinct, to be arranged in order of descending modulus. An arbitrary vector $\mathbf{y}_0$ is taken and two sequences of vectors $\mathbf{y}_i$ and $\mathbf{z}_i$ are formed from the relations
$$\mathbf{z}_{i+1} = \mathbf{A}\mathbf{y}_i, \tag{23}$$
$$\mathbf{y}_{i+1} = \mathbf{z}_{i+1} \div (\text{numerically largest element of } \mathbf{z}_{i+1}). \tag{24}$$
The vectors $\mathbf{y}_i$ form a sequence, each member of which has its largest element equal to unity. If we assume that the $\mathbf{x}_i$ are normalized so that their numerically largest elements are unity, then $\mathbf{y}_i$ tends to the vector $\mathbf{x}_1$ corresponding to the root λ_1. For $\mathbf{y}_0$ may be expressed in terms of the latent vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ by the relation
$$\mathbf{y}_0 = \sum_1^n \alpha_i \mathbf{x}_i. \tag{25}$$

If C_k is a constant corresponding to division by the largest element of the vector, we have

$$C_k \mathbf{y}_k = \sum_1^n \alpha_i \lambda_i^k \mathbf{x}_i = \lambda_1^k \left\{ \alpha_1 \mathbf{x}_1 + \sum_2^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i \right\}. \tag{26}$$

Now $(\lambda_i/\lambda_1)^k \to 0$ as $k \to \infty$ $(i \neq 1)$, and hence

$$C_k \mathbf{y}_k \to \lambda_1^k \alpha_1 \mathbf{x}_1, \tag{27}$$

which means that $\mathbf{y}_k$ tends to $\mathbf{x}_1$, since C_k is chosen so that the largest component is unity. The speed at which the convergence takes place depends on the ratio λ_2/λ_1; if this is almost equal in modulus to unity convergence may be slow.

6. The rate of convergence may often be improved by a very simple device. If $\mathbf{x}_i$ is a latent vector of $\mathbf{A}$ then we have

$$(\mathbf{A} - p\mathbf{I})\mathbf{x}_i = (\lambda_i - p)\mathbf{x}_i, \tag{28}$$

for any value of p. The latent vectors of $\mathbf{A} - p\mathbf{I}$ are therefore the same as those of $\mathbf{A}$ and the latent roots differ from those of $\mathbf{A}$ by the quantity p.

Suppose the latent roots of $\mathbf{A}$ to be 6, 5, 4, 3, 2, 1. If we iterate with $\mathbf{A}$ we tend to $\mathbf{x}_1$ at a speed determined by the rate at which $(5/6)^k \to 0$. If we iterate with $\mathbf{A} - 3\mathbf{I}$ we tend to $\mathbf{x}_1$ at a speed determined by the rate at which $(2/3)^k \to 0$, which is more than twice as great. If we iterate with $(\mathbf{A} - 4\mathbf{I})$, whose roots are $2, 1, 0, -1, -2, -3$, we tend to $\mathbf{x}_6$ at a speed determined by the rate at which $(2/3)^k \to 0$.

7. The process may be used to give convergence either to the largest or to the smallest root; it cannot be used to determine the intermediate roots. These can be found by a process of successive root-removal described later in §§ 10–13. Sometimes, however, the root nearest to a given value is required. To find this we may make use of a device similar to that of § 6.

From equation (28), with $\mathbf{x}_i$ replaced by $\mathbf{x}$, λ_i by λ, we obtain the equation

$$(\mathbf{A} - p\mathbf{I})^{-1}\mathbf{x} = (\lambda - p)^{-1}\mathbf{x}. \tag{29}$$

This shows that the latent vectors of $(\mathbf{A} - p\mathbf{I})^{-1}$ are the same as those of $\mathbf{A}$, but the latent root corresponding to λ_i is $(\lambda_i - p)^{-1}$. The dominant root of $(\mathbf{A} - p\mathbf{I})^{-1}$ will correspond to the root λ_i of $\mathbf{A}$ nearest to p because this will give the greatest value of $(\lambda - p)^{-1}$. It is unnecessary to compute the inverse of $\mathbf{A} - p\mathbf{I}$ explicitly; indeed, it would be uneconomical to do so in general. We perform Gaussian elimination, or triangular decomposition with interchanges, on the matrix $\mathbf{A} - p\mathbf{I}$. Iteration may then be carried out by using the relations

$$(\mathbf{A} - p\mathbf{I})\mathbf{z}_{i+1} = \mathbf{y}_i, \tag{30}$$

$$\mathbf{y}_{i+1} = \mathbf{z}_{i+1} \div (\text{numerically largest element of } \mathbf{z}_{i+1}), \tag{31}$$

as described in Chapter 2, §§ 9–11. If p is close to a latent root, the matrix $\mathbf{A} - p\mathbf{I}$ will be ill-conditioned. Nevertheless, the accuracy of the latent vectors determined in this way is unaffected; see [**32**].

8. A second iterative method for finding the latent root of largest modulus and its corresponding vector is that of matrix powering. If the sequence $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^4, \mathbf{A}^8, \ldots$ is formed, then all the columns of $\mathbf{A}^{2^k}$ become parallel to the dominant latent vector. This can be seen if we express $\mathbf{A}$ in terms of its columns in the form

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n].$$

Then
$$\left.\begin{aligned} \mathbf{A}^2 &= \mathbf{A}\mathbf{A} = [\mathbf{A}\mathbf{a}_1, \mathbf{A}\mathbf{a}_2, \ldots, \mathbf{A}\mathbf{a}_n], \\ \mathbf{A}^4 &= \mathbf{A}^2\mathbf{A}^2 = [\mathbf{A}^3\mathbf{a}_1, \mathbf{A}^3\mathbf{a}_2, \ldots, \mathbf{A}^3\mathbf{a}_n], \\ \mathbf{A}^{2^k} &= [\mathbf{A}^{2^k-1}\mathbf{a}_1, \mathbf{A}^{2^k-1}\mathbf{a}_2, \ldots, \mathbf{A}^{2^k-1}\mathbf{a}_n]. \end{aligned}\right\} \tag{32}$$

The proofs of § 5 show that each of the columns in parentheses is ultimately parallel to the dominant latent vector. The speed of convergence is given by the rate at which $(\lambda_2/\lambda_1)^{2^k} \to 0$. This gives more rapid convergence than the previous iterative method, but each iteration requires n^3 multiplications instead of n^2. It is readily seen that for a given ratio of λ_2 and λ_1, matrix powering is the more efficient method only for matrices of low order. The method has the further disadvantage that if $\mathbf{A}$ contains a number of zero elements, then these zeros do not persist in the matrix powers.

METHODS FOR FINDING A SUBDOMINANT ROOT

9. The iterative methods described are suitable for finding the greatest or smallest root of a matrix. To find a subdominant root by these methods the dominant root must first be removed from the matrix.

For a symmetric matrix the simplest method is the following. Suppose we have the root λ_1 and corresponding vector $\mathbf{x}_1$, normalized so that $\mathbf{x}_1'\,\mathbf{x}_1 = 1$, of a matrix $\mathbf{A}$. Then the matrix $\mathbf{A}_1$, defined by

$$\mathbf{A}_1 = \mathbf{A} - \lambda_1\mathbf{x}_1\mathbf{x}_1', \tag{33}$$

has the same latent roots and vectors as $\mathbf{A}$ except that the root corresponding to λ_1 has become zero. For we have

$$\mathbf{A}_1\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1 - \lambda_1\mathbf{x}_1\mathbf{x}_1'\mathbf{x}_1 = \lambda_1\mathbf{x}_1 - \lambda_1\mathbf{x}_1 = 0.$$

Also, if λ_2 and $\mathbf{x}_2$ are another root and vector of $\mathbf{A}$, then

$$\mathbf{A}_1\mathbf{x}_2 = \mathbf{A}\mathbf{x}_2 - \lambda_1\mathbf{x}_1\mathbf{x}_1'\mathbf{x}_2 = \lambda_2\mathbf{x}_2,$$

since $\mathbf{x}_1'\mathbf{x}_2 = 0$ by the orthogonality relation for latent vectors of a symmetric matrix.

10. For unsymmetric matrices the following is probably the simplest method of removing a known root λ_1 and vector $\mathbf{x}_1$. The matrix $\mathbf{A}$ is written in partitioned form as

$$\mathbf{A} = \begin{bmatrix} \mathbf{p}_1' \\ \cdots \\ \mathbf{B} \end{bmatrix}, \tag{34}$$

where $\mathbf{p}_1'$ is the first row of $\mathbf{A}$. The known vector $\mathbf{x}_1$ is normalized so that its first component is unity. The matrix $\mathbf{A}_1$ is then computed from the relation

$$\mathbf{A}_1 = \mathbf{A}-\mathbf{x}_1\mathbf{p}_1'. \tag{35}$$

If λ_2 and $\mathbf{x}_2$ are latent roots and vectors of $\mathbf{A}$, $\mathbf{x}_2$ being normalized so that its first component is unity, then $\mathbf{x}_1-\mathbf{x}_2$ is a latent vector of $\mathbf{A}_1$, with latent root λ_2. For

$$\begin{aligned}\mathbf{A}_1(\mathbf{x}_1-\mathbf{x}_2) &= \mathbf{A}(\mathbf{x}_1-\mathbf{x}_2)-\mathbf{x}_1\mathbf{p}_1'(\mathbf{x}_1-\mathbf{x}_2)\\ &= \lambda_1\mathbf{x}_1-\lambda_2\mathbf{x}_2-\mathbf{x}_1(\lambda_1-\lambda_2)\\ &= \lambda_2(\mathbf{x}_1-\mathbf{x}_2).\end{aligned} \tag{36}$$

Hence the latent roots of $\mathbf{A}_1$ are the same as those of $\mathbf{A}$, except that the root λ_1 has become zero, since $\mathbf{A}_1\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1-\mathbf{x}_1\mathbf{p}_1'\mathbf{x}_1 = \lambda_1\mathbf{x}_1-\mathbf{x}_1\lambda_1 = 0$. The latent vectors are simply related to those of $\mathbf{A}$. It is easily seen that the first row of $\mathbf{A}_1$ is zero throughout and that, since each of the latent vectors $\mathbf{x}_1-\mathbf{x}_i$ of $\mathbf{A}_1$ has a zero component in its first position, we need work only with the matrix of order $n-1$ in the bottom right-hand corner of $\mathbf{A}_1$. Hence the order of the relevant matrix is reduced by unity with each root we find.

For the computation of the required vector, we first find a latent vector $\mathbf{y}_2$ of the $(n-1)\times(n-1)$ matrix obtained from $\mathbf{A}_1$, and extend it to order n by giving it a zero first component. Then the latent vector $\mathbf{x}_2$ of $\mathbf{A}$, corresponding to $\mathbf{y}_2$, is given by

$$\mathbf{x}_2 = \mathbf{x}_1+k\mathbf{y}_2, \tag{37}$$

where we use the extended $\mathbf{y}_2$. The factor k is necessary because of a normalizing factor in $\mathbf{y}_2$. Multiplying (37) by $\mathbf{p}_1'$, we have

$$\lambda_2 = \lambda_1+k\mathbf{p}_1'\mathbf{y}_2, \tag{38}$$

and from this we obtain k. Equation (37) then provides the required vector $\mathbf{x}_2$.

In this description it was assumed that $\mathbf{x}_1$ had been normalized so that its first element is unity. From the point of view of numerical convenience it is better to normalize so that the largest element is unity. The analysis is unaltered, but since we then use the row of $\mathbf{A}$ in the position corresponding to this largest element instead of the first row, the notation is not so convenient.

11. A simple example will illustrate the method of root-removal. A root and vector of the matrix

$$\begin{bmatrix} 2 & 3 & 2\\ 10 & 3 & 4\\ 3 & 6 & 1\end{bmatrix} \text{ are respectively } \lambda = -2 \text{ and } \mathbf{x} = \begin{bmatrix} 1\\ 2\\ -5\end{bmatrix}.$$

Normalizing $\mathbf{x}$ so that its largest component is 1, we obtain

$$\mathbf{x} = \begin{bmatrix} -0{\cdot}2\\ -0{\cdot}4\\ 1{\cdot}0\end{bmatrix}.$$

3

The last row of **A** is therefore used in the root-removal process and we find

$$\mathbf{A}_1 = \begin{bmatrix} 2{\cdot}6 & 4{\cdot}2 & 2{\cdot}2 \\ 11{\cdot}2 & 5{\cdot}4 & 4{\cdot}4 \\ 0 & 0 & 0 \end{bmatrix}.$$

It is clear that the latent vectors of $\mathbf{A}_1$ all have zero for their last component and therefore we need only work with vectors of order 2 and the matrix

$$\begin{bmatrix} 2{\cdot}6 & 4{\cdot}2 \\ 11{\cdot}2 & 5{\cdot}4 \end{bmatrix}$$

of order 2. A latent vector of this matrix is the vector $\begin{bmatrix} -3 \\ 4 \end{bmatrix}$, with the corresponding latent root equal to -3.

We can then remove this vector of order 2 from the matrix of order 2. To do this we write the vector in the form $\begin{bmatrix} -0{\cdot}75 \\ 1{\cdot}00 \end{bmatrix}$, and use the last row in the root-removal process. This gives the matrix

$$\begin{bmatrix} 11 & 8{\cdot}25 \\ 0 & 0 \end{bmatrix}.$$

It is clear that the two latent vectors of this matrix have zero in the last position and that we need only work with the single element 11. The last latent root is therefore 11.

We have thus obtained 3 latent vectors,

$$\begin{bmatrix} -0{\cdot}2 \\ -0{\cdot}4 \\ 1{\cdot}0 \end{bmatrix}, \quad \begin{bmatrix} -0{\cdot}75 \\ 1{\cdot}00 \end{bmatrix} \quad \text{and} \quad [1{\cdot}0],$$

of which only the first is a latent vector of the original matrix **A**.

12. For a matrix of order n we would have found n vectors of the form:

one vector $\mathbf{u}_1$ of order n (a true latent vector of the matrix **A**),

one vector $\mathbf{u}_2$ of order $n-1$,

one vector $\mathbf{u}_3$ of order $n-2$,

....

one vector $\mathbf{u}_n$ of order 1.

To obtain from the vector of order $n-r$ a true latent vector of the original matrix **A** we would proceed in the following r steps.

A vector of order $n-r+1$ would be computed, using the vector $\mathbf{u}_r$ of order $n-r+1$, from relations of types (37) and (38).

From this vector one of order $n-r+2$ would be computed, using these relations again with the latent vector $\mathbf{u}_{r-1}$ of order $n-r+2$, and so on, until we obtained a true latent vector of order n.

13. In the above example the vector $\begin{bmatrix} -0{\cdot}75 \\ 1{\cdot}00 \end{bmatrix}$ leads to a true latent vector $\mathbf{x}_2$ from the relation

$$\mathbf{x}_2 = \begin{bmatrix} -0{\cdot}2 \\ -0{\cdot}4 \\ 1{\cdot}0 \end{bmatrix} + k \begin{bmatrix} -0{\cdot}75 \\ 1{\cdot}00 \\ 0{\cdot}00 \end{bmatrix}.$$

Multiplying this by [3, 6, 1], the last row of **A**, we find, from (38), the result $-3 = -2 + 3{\cdot}75k$, so that $k = -\frac{4}{15}$, and then

$$\mathbf{x}_2 = \begin{bmatrix} 0 \\ -\frac{2}{3} \\ 1 \end{bmatrix}.$$

The third vector may be found in two similar steps, given below:

(i)
$$\mathbf{y}_3 = \begin{bmatrix} -0{\cdot}75 \\ 1{\cdot}00 \end{bmatrix} + k \begin{bmatrix} 1{\cdot}00 \\ 0{\cdot}00 \end{bmatrix}.$$

Multiplication by [11·2, 5·4] gives

$$11 = -3 + 11{\cdot}2k, \text{ so that } k = 1{\cdot}25, \text{ and } \mathbf{y}_3 = \begin{bmatrix} 0{\cdot}50 \\ 1{\cdot}00 \end{bmatrix}.$$

(ii)
$$\mathbf{x}_3 = \begin{bmatrix} -0{\cdot}2 \\ -0{\cdot}4 \\ 1{\cdot}0 \end{bmatrix} + k \begin{bmatrix} 0{\cdot}5 \\ 1{\cdot}0 \\ 0{\cdot}0 \end{bmatrix}.$$

Multiplication by [3, 6, 1] gives

$$11 = -2 + 7{\cdot}5k, \quad \text{so that} \quad k = \tfrac{26}{15}, \quad \text{and}$$

$$\mathbf{x}_3 = \begin{bmatrix} \frac{2}{3} \\ \frac{4}{3} \\ 1 \end{bmatrix}.$$

14. We have described these iterative methods in some detail because they are simple and, besides being of value in their own right, they are frequently used in conjunction with other methods. For symmetric matrices there are a number of more powerful methods which would, in most circumstances, be used instead. We now describe some of these.

THE METHOD OF JACOBI

15. This method depends upon the fact that if $\mathbf{TT}' = \mathbf{I}$, that is, if **T** is *orthogonal*, then the roots of $\mathbf{TAT}'$ are the same as those of **A**. For

$$\mathbf{T}(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{T}' = \mathbf{TAT}' - \lambda\mathbf{I}, \qquad (39)$$

so that the zeros of $|\mathbf{A} - \lambda\mathbf{I}|$ are the same as those of $|\mathbf{TAT}' - \lambda\mathbf{I}|$.

A simple orthogonal matrix **T** is given by $T_{ii} = 1 (i \neq p, q)$, $T_{ij} = 0$ for all other i, j except that

$$T_{pp} = \cos\theta, \quad T_{qq} = \cos\theta, \quad T_{pq} = \sin\theta, \quad T_{qp} = -\sin\theta. \tag{40}$$

If we form **TAT′**, only the p and q rows and the p and q columns of **A** are altered, and the matrix remains symmetric. The (p, q) element of **TAT′** is given by $a_{pq}\cos 2\theta - \frac{1}{2}(a_{pp} - a_{qq})\sin 2\theta$, from which it follows that if

$$\tan 2\theta = 2a_{pq}/(a_{pp} - a_{qq}), \tag{41}$$

then the (p, q) element of **TAT′** is zero.

It is easy to prove that both the sum of the diagonal elements (the *trace*) and the sum of the squares of the off-diagonal elements other than the (p, q) and (q, p) elements are unchanged. If we choose a succession of **T** matrices and successively premultiply by **T** and postmultiply by **T′**, each **T** matrix being chosen to make zero the largest off-diagonal elements in the resulting matrix at that stage, then we ultimately obtain a diagonal matrix. We have

$$\mathbf{T}_r \mathbf{T}_{r-1} \dots \mathbf{T}_1 \mathbf{A} \mathbf{T}_1' \mathbf{T}_2' \dots \mathbf{T}_r' = \mathbf{D}, \tag{42}$$

where **D** is a diagonal matrix. Since the product of orthogonal matrices is itself orthogonal, the elements of **D** are the latent roots of **A**. If the latent vectors of **A** are wanted they are given by the columns of the product matrix

$$\mathbf{T}_1' \mathbf{T}_2' \dots \mathbf{T}_r' = \mathbf{S}. \tag{43}$$

This follows from (42), since

$$\mathbf{S}'\mathbf{A}\mathbf{S} = \mathbf{D}, \tag{44}$$

giving

$$\mathbf{A}\mathbf{S} = (\mathbf{S}')^{-1}\mathbf{D} = \mathbf{S}\mathbf{D}. \tag{45}$$

GIVENS' METHOD

16. Each step of the reduction in Jacobi's method, consisting of premultiplication by the matrix **T**, and postmultiplication by its transpose **T′**, where the non-zero elements of **T** are given by (40), is called a *rotation.* The pair of values of p, q is called the *plane* of the rotation, and θ the *angle* of the rotation.

The method of Givens is similar to that of Jacobi, inasmuch as each step of it consists of a rotation. In this case, however, we choose θ so that the element in the $(p-1, q)$ position $(p > 1)$ becomes zero. This gives

$$\tan\theta = a_{p-1,q}/a_{p-1,p}. \tag{46}$$

The rotations are applied systematically to the matrix to make zero, in order, the following elements:

	Positions of zero elements	*Planes of rotation*
1st row	$(1,3), (1,4), (1,5), \dots, (1,n)$	$(2,3), (2,4), (2,5), \dots, (2,n)$
2nd row	$(2,4), (2,5), \dots, (2,n)$	$(3,4), (3,5), \dots, (3,n)$
	....	....
$(n-2)$th row	$(n-2, n)$	$(n-1, n)$

In contrast to the Jacobi rotations, each zero element, once produced, persists throughout the subsequent transformations. The symmetry of the matrix is preserved, so that after carrying out the above $\frac{1}{2}(n-1)(n-2)$ rotations, all elements are zero, other than those in the principal diagonal and the immediately adjacent diagonals, one on either side. The matrix is then said to be of *triple-diagonal* form, or sometimes *tri-diagonal* or *co-diagonal* form.

As the work progresses the amount of computation in a rotation becomes steadily less. At the stage when zeros are introduced in the rth row, we are effectively working with a matrix of order $n-r+1$ since the first $r-1$ rows and columns remain unaltered. Because of this and the non-iterative nature of the computation, the reduction to triple-diagonal form takes about one-twentieth of the time for the reduction to diagonal form by Jacobi's method [**216**].

17. The latent roots of the symmetric triple-diagonal form are the same as those of the original matrix, and we now consider their evaluation. Since other methods lead to triple-diagonal matrices and also many latent root problems give rise to matrices which are already in this form, the solution of this problem is important quite apart from its present context.

DETERMINATION OF THE LATENT ROOTS AND VECTORS OF A SYMMETRIC TRIPLE-DIAGONAL MATRIX

18. The method we describe, sometimes called the *method of bisections*, is often much slower than alternative methods in existence, but it is comparatively simple, and has such remarkable numerical stability that it is frequently used. It depends on the following result [**30, 31**].

Let $p_r(\lambda)$ be the value of the rth leading principal minor of $\mathbf{C}-\lambda\mathbf{I}$, where $\mathbf{C}$ is a symmetric triple-diagonal matrix, and $p_0(\lambda) = 1$. Then the number, $s(\lambda)$, of agreements in sign between consecutive members of the sequence $p_0(\lambda), p_1(\lambda), p_2(\lambda), \ldots, p_n(\lambda)$ is equal to the number of latent roots of $\mathbf{C}$ which are greater than λ. (Note that $s(\lambda)$ is an integer between 0 and n inclusive.)

Let the non-zero elements of $\mathbf{C}$ be denoted by

$$d_{r,r} = \alpha_r, \qquad d_{r-1,r} = d_{r,r-1} = \beta_r. \tag{47}$$

We assume that no β_r vanishes since otherwise the problem could be broken up into the solution of a number of smaller triple-diagonal matrices. With this assumption, it may be shown that $\mathbf{C}$ has no multiple latent roots. The values of the leading principal minors, $p_r(\lambda)$, may be computed from the recurrence relations,

$$\left.\begin{aligned} &p_0(\lambda) = 1, \qquad p_1(\lambda) = \alpha_1 - \lambda, \\ &p_r(\lambda) = (\alpha_r - \lambda)p_{r-1}(\lambda) - \beta_r^2 p_{r-2}(\lambda) \qquad (r = 2, 3, \ldots, n). \end{aligned}\right\} \tag{48}$$

If any p_r is zero, its sign should be regarded as the opposite of that of p_{r-1}. With the assumption $\beta_r \neq 0$ we cannot have two consecutive p_r which are zero.

19. The result may be applied to determine any latent root, the kth, λ_k say, as follows. Suppose it is known that

$$s(a) \geqslant k, \qquad s(b) < k,$$

then $a < \lambda_k \leqslant b$. Clearly, if $s\{\frac{1}{2}(a+b)\} < k$, then λ_k lies between a and $\frac{1}{2}(a+b)$, while if $s\{\frac{1}{2}(a+b)\} \geqslant k$, λ_k lies between $\frac{1}{2}(a+b)$ and b. In either case, an evaluation of $s(\lambda)$ at the mid-point of the interval (a, b) enables us to locate it in an interval of width $\frac{1}{2}(b-a)$. By making t successive applications of this principle we locate λ_k in an interval of width $2^{-t}(b-a)$; this requires the evaluation of $s(\lambda)$ at t points.

In order to begin the process we need values for a and b. Now if λ is any latent root of $\mathbf{C}$ and $\mathbf{x}$ is the corresponding latent vector, we have

$$\beta_r x_{r-1} + \alpha_r x_r + \beta_{r+1} x_{r+1} = \lambda x_r, \tag{49}$$

where β_1 and β_{n+1} are to be taken as zero. Hence

$$\lambda = \beta_r \frac{x_{r-1}}{x_r} + \alpha_r + \beta_{r+1} \frac{x_{r+1}}{x_r}. \tag{50}$$

Taking x_r to be the element of $\mathbf{x}$ of greatest modulus, we deduce that

$$|\lambda| \leqslant |\beta_r| + |\alpha_r| + |\beta_{r+1}|. \tag{51}$$

Adequate values of a and b are therefore given by

$$a, b = \mp \max\{|\beta_r| + |\alpha_r| + |\beta_{r+1}|\}. \tag{52}$$

20. The method has the following advantages:

(i) We may find a latent root of any prescribed enumeration without determining the others. As a corollary, we are not obliged to find all the roots to the same precision; the number of steps t may be pre-assigned for each root.

(ii) The time taken to find each root is directly proportional to n (*not* to a higher power of n), and to the number of steps, i.e. the required precision. It is independent of the separation of the roots.

(iii) If all the elements α_r and β_r are numerically less than $\frac{1}{4}$, then t steps of floating-point computation with t binary digits in the mantissa yields the value of any latent root with an error not exceeding 4×2^{-t}.

The proof of (iii) requires a detailed error analysis and we refer the reader to [**32**].

21. The latent vectors of the original matrix $\mathbf{A}$ are equal to those of the triple-diagonal matrix $\mathbf{C}$, premultiplied by the product $\mathbf{S}$ of the rotation matrices; compare § 15. To complete the solution of the problem we therefore have to find the latent vectors of $\mathbf{C}$. Since the latent roots of $\mathbf{C}$ are readily computable to high accuracy, we might expect that this would be a comparatively trivial problem. Formally, the components of the latent vector corresponding to λ_k are given by

$$p_0(\lambda_k), \quad -p_1(\lambda_k)/\beta_2, \quad p_2(\lambda_k)/(\beta_2\beta_3), \ldots, \quad (-)^{n-1}p_{n-1}(\lambda_k)/(\beta_2\beta_3\cdots\beta_n). \tag{53}$$

Now although the $p_r(\lambda)$ may be used to determine the latent roots in a very stable manner, the determination of the latent vector from (53) is unstable. Even if the n quantities are determined *exactly* for a value of λ_k which is itself almost exactly a latent root, the resulting vector may be very nearly orthogonal to the true latent vector.

Accurate vectors may be found by the following process. Corresponding to each computed λ_k, the matrix $\mathbf{C}-\lambda_k\mathbf{I}$ is reduced to an upper triangular matrix $\mathbf{T}_k$, by Gaussian elimination with interchanges. The equations $\mathbf{T}_k\mathbf{x} = \mathbf{e}$, where $\mathbf{e}$ is the vector with all its components equal to unity, are then solved, and the solution $\mathbf{x}$ is the latent vector corresponding to λ_k. For further details the reader is referred to [**22**].

HOUSEHOLDER'S METHOD

22. For several years Givens' method was probably the best of the known methods for symmetric matrices of general form. Recently, Householder [**28**] has suggested an alternative method of reduction to triple-diagonal form using elementary orthogonal transformations which are not plane rotations. This method is a substantial improvement on Givens' method. It requires only half as many arithmetical operations, has an even smaller maximum rounding error and, if the vectors are wanted, requires less storage. For details of the practical application of the method, see [**25**].

UNSYMMETRIC MATRICES

23. There are no methods for unsymmetric matrices which are as satisfactory as those we have just described for symmetric matrices. However, a number of direct methods exist which, in general, have advantages over the iterative methods of §§ 5–13 when all the roots are required. Because of the inherent instability of the unsymmetric problem, considerable attention to arithmetical detail is necessary in designing effective programmes. The presentation of these methods is beyond the scope of this manual, but references to the more satisfactory ones are given in the Bibliography on pages 147–148.

4

LINEAR EQUATIONS AND MATRICES: ITERATIVE METHODS

DEFINITIONS

1. In Chapters 1 and 2, some *direct* methods have been described; these yield solutions after an amount of computation that can be specified in advance. In contrast, the *iterative* or *indirect* methods of this chapter start from an approximation to the true solution and, if successful, derive a convergent sequence of closer approximations from a computational cycle repeated as often as may be necessary for the purpose. This means that in a direct method the number of arithmetic operations is independent of the accuracy required in the solution (provided the word length of our computer is adequate to offset any ill-conditioning), while in an iterative process the amount of arithmetic depends upon the accuracy required.

When a genuine choice is available, usually the direct methods should be preferred, but for matrices containing a large proportion of zero elements, such as arise in the solution of partial differential equations, iterative methods which preserve these elements, and therefore involve a smaller amount of machine store, can be advantageous.

For desk-machine work the 'relaxation' methods developed by Southwell, Fox and others are very suitable, and descriptions are available in the previous edition of this manual and elsewhere [**103**]. However, these methods are not convenient for use on automatic computers, and we shall not discuss them here.

2. We proceed to describe two basic iterative methods which can be applied to the solution of a general set of linear equations

$$\mathbf{Ax} = \mathbf{b}. \tag{1}$$

It is convenient first to reduce (1) to the form

$$(\mathbf{I}-\mathbf{L}-\mathbf{U})\mathbf{x} = \mathbf{d}, \tag{2}$$

where **L** and **U** are respectively lower and upper triangular matrices with null diagonals, and **I** is the unit matrix. This is achieved by rearranging the equations so that no diagonal coefficient vanishes, and then dividing each equation by the corresponding diagonal coefficient. Furthermore it is advantageous, whenever possible, similarly to manoeuvre the largest coefficients into the diagonal positions, since the iterations are likely then to converge more rapidly.

3. To illustrate the methods presented we shall consider the equations

$$\left.\begin{aligned} x_1-\tfrac{1}{4}x_2-\tfrac{1}{4}x_3 \qquad &= \tfrac{1}{2}, \\ -\tfrac{1}{4}x_1+x_2 \qquad -\tfrac{1}{4}x_4 &= \tfrac{1}{2}, \\ -\tfrac{1}{4}x_1 \qquad +x_3-\tfrac{1}{4}x_4 &= \tfrac{1}{4}, \\ -\tfrac{1}{4}x_2-\tfrac{1}{4}x_3+x_4 &= \tfrac{1}{4}. \end{aligned}\right\} \tag{3}$$

(It is shown in Chapter 12 that equations of this form arise in the solution of Laplace's equation.)

These equations have the form (2) with

$$\mathbf{U}=\begin{bmatrix} 0 & \tfrac{1}{4} & \tfrac{1}{4} & 0 \\ & 0 & 0 & \tfrac{1}{4} \\ & & 0 & \tfrac{1}{4} \\ & & & 0 \end{bmatrix}, \qquad \mathbf{L}=\mathbf{U}'. \tag{4}$$

Their exact solution is $x_1 = 0{\cdot}875 = x_2$, $x_3 = 0{\cdot}625 = x_4$.

4. We shall denote the nth approximation to the solution vector $\mathbf{x}$ by $\mathbf{x}^{(n)}$ and the error vector $\mathbf{x}-\mathbf{x}^{(n)}$ by $\mathbf{e}^{(n)}$. Obviously $\mathbf{e}^{(n)}$ cannot be evaluated before the solution is available and so, to indicate the progress of the computations, we examine the elements of the *displacement vector* $\boldsymbol{\gamma}^{(n)}$, defined by

$$\boldsymbol{\gamma}^{(n)} = \mathbf{x}^{(n+1)}-\mathbf{x}^{(n)}. \tag{5}$$

It follows from repeated application of (5) that

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(0)}+\sum_{i=0}^{n}\boldsymbol{\gamma}^{(i)}. \tag{6}$$

Hence, for a process to converge, it is necessary that $\boldsymbol{\gamma}^{(n)}\to 0$ in such a way that the series (6) converges as $n\to\infty$.

SIMPLE OR JACOBI ITERATION

5. In this process, applied to equations (3), we calculate $\mathbf{x}^{(n+1)}$ from $\mathbf{x}^{(n)}$ according to the formulae

$$\left.\begin{aligned} x_1^{(n+1)} &= \qquad \tfrac{1}{4}x_2^{(n)}+\tfrac{1}{4}x_3^{(n)} \qquad +\tfrac{1}{2}, \\ x_2^{(n+1)} &= \tfrac{1}{4}x_1^{(n)} \qquad\qquad +\tfrac{1}{4}x_4^{(n)}+\tfrac{1}{2}, \\ x_3^{(n+1)} &= \tfrac{1}{4}x_1^{(n)} \qquad\qquad +\tfrac{1}{4}x_4^{(n)}+\tfrac{1}{4}, \\ x_4^{(n+1)} &= \qquad \tfrac{1}{4}x_2^{(n)}+\tfrac{1}{4}x_3^{(n)} \qquad +\tfrac{1}{4}. \end{aligned}\right\} \tag{7}$$

This is called a *method of simultaneous displacements* since no element of $\mathbf{x}^{(n+1)}$ is used in this iteration until every element has been calculated, and then $\mathbf{x}^{(n+1)}$ replaces $\mathbf{x}^{(n)}$ entirely for the next cycle.

For the general case (2) we use

$$\mathbf{x}^{(n+1)} = (\mathbf{L}+\mathbf{U})\,\mathbf{x}^{(n)}+\mathbf{d}. \tag{8}$$

From (5) and (8), the change to be made to $\mathbf{x}^{(n)}$ is

$$\boldsymbol{\gamma}^{(n)} = \mathbf{d}-(\mathbf{I}-\mathbf{L}-\mathbf{U})\,\mathbf{x}^{(n)}. \tag{9}$$

If we take the initial values $\mathbf{x}^{(0)}$ to be zero in (7), we obtain the following sequence of approximations and displacements for equations (3).

$\mathbf{x}^{(1)}$	$\gamma^{(1)}$	$\mathbf{x}^{(2)}$	$\gamma^{(2)}$	$\mathbf{x}^{(3)}$	$\gamma^{(3)}$	$\mathbf{x}^{(4)}$	$\gamma^{(4)}$	$\mathbf{x}^{(5)}$
0·5	0·1875	0·6875	0·09375	0·78125	0·04687	0·82812	0·02344	0·85156
0·5	0·1875	0·6875	0·09375	0·78125	0·04687	0·82812	0·02344	0·85156
0·25	0·1875	0·4375	0·09375	0·53125	0·04687	0·57812	0·02344	0·60156
0·25	0·1875	0·4375	0·09375	0·53125	0·04687	0·57812	0·02344	0·60156

We observe that $\gamma^{(n)}$ is halved by each iteration after the first, so that five-decimal accuracy will be achieved after seventeen cycles.

GAUSS-SEIDEL OR LIEBMANN ITERATION

6. For the equations (3), the elements of $\mathbf{x}^{(n+1)}$ are determined in succession from the equations

$$\left.\begin{aligned} x_1^{(n+1)} &= \tfrac{1}{4}x_2^{(n)} + \tfrac{1}{4}x_3^{(n)} + \tfrac{1}{2}, \\ x_2^{(n+1)} &= \tfrac{1}{4}x_1^{(n+1)} + \tfrac{1}{4}x_4^{(n)} + \tfrac{1}{2}, \\ x_3^{(n+1)} &= \tfrac{1}{4}x_1^{(n+1)} + \tfrac{1}{4}x_4^{(n)} + \tfrac{1}{4}, \\ x_4^{(n+1)} &= \tfrac{1}{4}x_2^{(n+1)} + \tfrac{1}{4}x_3^{(n+1)} + \tfrac{1}{4}. \end{aligned}\right\} \tag{10}$$

Corresponding elements of $\mathbf{x}^{(n+1)}$ now replace those of $\mathbf{x}^{(n)}$ in the calculation as soon as they have been computed, and so this is called a *method of successive displacements*. A complete iteration cycle comprises one such displacement for each equation.

The matrix expression of this process for the general case (2) is

$$\mathbf{x}^{(n+1)} = \mathbf{L}\mathbf{x}^{(n+1)} + \mathbf{U}\mathbf{x}^{(n)} + \mathbf{d}. \tag{11}$$

From (5) and (11) it follows that the change made to $\mathbf{x}^{(n)}$ is

$$\gamma^{(n)} = \mathbf{d} + \mathbf{L}\mathbf{x}^{(n+1)} - (\mathbf{I} - \mathbf{U})\,\mathbf{x}^{(n)}. \tag{12}$$

For equations (3), if we take the initial values $\mathbf{x}^{(0)}$ to be zero, the following sequence of approximations and displacements is obtained from (10).

$\mathbf{x}^{(1)}$	$\gamma^{(1)}$	$\mathbf{x}^{(2)}$	$\gamma^{(2)}$	$\mathbf{x}^{(3)}$	$\gamma^{(3)}$	$\mathbf{x}^{(4)}$	$\gamma^{(4)}$	$\mathbf{x}^{(5)}$
0·5	0·25	0·75	0·09375	0·84375	0·02344	0·86719	0·00586	0·87305
0·625	0·1875	0·8125	0·04688	0·85938	0·01172	0·87110	0·00292	0·87402
0·375	0·1875	0·5625	0·04688	0·60938	0·01172	0·62110	0·00292	0·62402
0·5	0·09375	0·59375	0·02344	0·61719	0·00586	0·62305	0·00146	0·62451

It is clear that $\gamma^{(n)}$ is multiplied by $\frac{1}{4}$ in each iteration after the second, so that five-decimal accuracy will be achieved after nine cycles.

CONVERGENCE OF THE ITERATIONS

7. A process converges if the corresponding sequence $\mathbf{e}^{(n)}$ tends to zero, where $\mathbf{e}^{(n)}$ denotes $\mathbf{x} - \mathbf{x}^{(n)}$. For the Jacobi iteration, (2) and (8) give

$$\mathbf{e}^{(n)} = (\mathbf{L} + \mathbf{U})\,\mathbf{e}^{(n-1)} = \ldots = (\mathbf{L} + \mathbf{U})^n\,\mathbf{e}^{(0)}. \tag{13}$$

For the Gauss–Seidel process, (2) and (11) lead to

$$\mathbf{e}^{(n)} = (\mathbf{I}-\mathbf{L})^{-1}\mathbf{U}\mathbf{e}^{(n-1)} = \ldots \doteq [(\mathbf{I}-\mathbf{L})^{-1}\mathbf{U}]^{n}\,\mathbf{e}^{(0)}. \qquad (14)$$

Now it may be shown, [7, § 2.06], [**216**], that any real symmetric matrix of order m has m linearly independent latent vectors $\mathbf{v}_i$, whether or not the corresponding latent roots λ_i are distinct. The matrix $\mathbf{L}+\mathbf{U}$ of (3), for example, has vectors

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix},$$

corresponding respectively to the latent roots $\frac{1}{2}, 0, 0, -\frac{1}{2}$.

Accordingly, for such an iteration matrix we can express the initial error vector $\mathbf{e}^{(0)}$ in the form

$$\mathbf{e}^{(0)} = \sum_{i=1}^{m} \alpha_i \mathbf{v}_i. \qquad (15)$$

Hence, for the Jacobi iteration,

$$\mathbf{e}^{(1)} = (\mathbf{L}+\mathbf{U})\,\mathbf{e}^{(0)} = \sum_{i=1}^{m} \alpha_i \lambda_i \mathbf{v}_i,$$

and

$$\mathbf{e}^{(n)} = \sum_{i=1}^{m} \alpha_i \lambda_i^{n} \mathbf{v}_i. \qquad (16)$$

For the iterations to converge from an arbitrary vector $\mathbf{x}^{(0)}$ it is clearly necessary that all the latent roots of the iteration matrix have modulus less than unity. The smaller the magnitude of the largest root, the faster the process converges.

8. In general, the iteration matrix will not be symmetric, and an unsymmetric matrix $\mathbf{B}$ of order m with repeated latent roots may have fewer than m independent latent vectors. For some repeated latent root λ of $\mathbf{B}$ there will then exist one or more *principal vectors* $\mathbf{w}$ *of grade* p $(p > 1)$ satisfying the equation

$$(\mathbf{B}-\lambda\mathbf{I})^{p-1}\mathbf{w} = \mathbf{v}, \qquad (17)$$

where $\mathbf{v}$ is a latent vector corresponding to λ. In such a case the latent vectors and principal vectors together comprise m linearly independent vectors.

As illustration, we have for equations (3),

$$(\mathbf{I}-\mathbf{L})^{-1}\mathbf{U} = \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & \frac{1}{16} & \frac{1}{16} & \frac{1}{4} \\ 0 & \frac{1}{16} & \frac{1}{16} & \frac{1}{4} \\ 0 & \frac{1}{32} & \frac{1}{32} & \frac{1}{8} \end{bmatrix}. \qquad (18)$$

This matrix is clearly unsymmetric, and it can be shown to possess three latent vectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ and a principal vector $\mathbf{w}$ of grade 2:

$$\mathbf{v}_1 = \begin{bmatrix} 4 \\ 2 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{w} = \begin{bmatrix} 0 \\ 0 \\ 4 \\ -1 \end{bmatrix},$$

and the associated latent roots are $\frac{1}{4}, 0, 0, 0$, respectively. It is readily verified that $(\mathbf{I}-\mathbf{L})^{-1}\mathbf{U}\mathbf{w} = \mathbf{v}_3$, and that $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ and $\mathbf{w}$ are linearly independent.

We now give the counterpart to (16) for an iteration matrix which has $m-1$ latent vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{m-1}$, and a principal vector of grade 2, $\mathbf{v}_m$, corresponding to the repeated latent root λ_{m-1}. In this case, it follows from (15) and (17) that

$$\begin{aligned} \mathbf{e}^{(1)} &= \sum_{i=1}^{m-1} \alpha_i \lambda_i \mathbf{v}_i + \alpha_m(\mathbf{v}_{m-1} + \lambda_{m-1}\mathbf{v}_m) \\ &= \sum_{i=1}^{m-2} \alpha_i \lambda_i \mathbf{v}_i + (\alpha_{m-1}\lambda_{m-1} + \alpha_m)\mathbf{v}_{m-1} + \alpha_m \lambda_{m-1}\mathbf{v}_m, \end{aligned} \tag{19}$$

and

$$\mathbf{e}^{(n)} = \sum_{i=1}^{m-2} \alpha_i \lambda_i^n \mathbf{v}_i + (\alpha_{m-1}\lambda_{m-1}^n + n\alpha_m \lambda_{m-1}^{n-1})\mathbf{v}_{m-1} + \alpha_m \lambda_{m-1}^n \mathbf{v}_m. \tag{20}$$

As before, it is necessary for convergence that all the latent roots have modulus less than unity. The association of the factor n with $\mathbf{v}_{m-1}$, as compared with (16), does not affect the convergence appreciably. For the Gauss–Seidel iteration (18), the associated root is zero and the term in $\mathbf{v}_{m-1}$ vanishes.

9. With arbitrary $\mathbf{L}$ and $\mathbf{U}$ it is difficult to guarantee in advance that the convergence condition will be satisfied, and theoretical results are available only for special classes of matrices. For example, if $\mathbf{I}-\mathbf{L}-\mathbf{U}$ is symmetric, then a necessary and sufficient condition for the Gauss–Seidel iteration to converge is that $\mathbf{I}-\mathbf{L}-\mathbf{U}$ should be positive definite (Chapter 3, § 3). Again, if none of the elements of $\mathbf{L}+\mathbf{U}$ is negative, then the Jacobi and Gauss–Seidel iterations either both converge or both diverge.

Another class of matrices which arises frequently in the study of partial differential equations consists of matrices possessing what is known as *Property A* [**112**], that is, the equations can be rearranged to provide $\mathbf{L}+\mathbf{U}$ with the form

$$\begin{bmatrix} \mathbf{O} & \mathbf{R} \\ \mathbf{Q} & \mathbf{O} \end{bmatrix}, \tag{21}$$

where $\mathbf{O}$ represents a null square submatrix. For this and certain other rearrangements of such equations the Gauss–Seidel iteration has latent roots equal to the squares of the latent roots of the Jacobi iteration (though the correspondence is not one-to-one), so that when the former converges it does so twice as fast as the latter.

These theorems are all illustrated by the equations (3), for which $\mathbf{I}-\mathbf{L}-\mathbf{U}$ is symmetric. This matrix has latent roots $1\frac{1}{2}, 1, 1, \frac{1}{2}$. Also it possesses Property A since the equations can be rewritten in the form

$$\left.\begin{aligned} x_1 \qquad -\tfrac{1}{4}x_2-\tfrac{1}{4}x_3 &= \tfrac{1}{2}, \\ x_4-\tfrac{1}{4}x_2-\tfrac{1}{4}x_3 &= \tfrac{1}{4}, \\ -\tfrac{1}{4}x_1-\tfrac{1}{4}x_4+\ x_2 \qquad &= \tfrac{1}{2}, \\ -\tfrac{1}{4}x_1-\tfrac{1}{4}x_4 \qquad +\ x_3 &= \tfrac{1}{4}. \end{aligned}\right\} \tag{22}$$

The largest latent root of $(\mathbf{I}-\mathbf{L})^{-1}\mathbf{U}$ is the square of that of $\mathbf{L}+\mathbf{U}$, and we have seen that the former iteration converges more rapidly.

SUCCESSIVE OVERRELAXATION

10. Much recent research has been directed to developing processes which converge more rapidly than those given in §§ 5, 6. Here we describe only the simplest of these iterations (but one of the most powerful), known variously as 'extrapolated Gauss–Seidel', 'extrapolated Liebmann', and 'successive overrelaxation'.

Successive overrelaxation is defined by the use of

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \omega[\mathbf{d}+\mathbf{L}\mathbf{x}^{(n+1)}-(\mathbf{I}-\mathbf{U})\,\mathbf{x}^{(n)}] \tag{23}$$

in place of (11). By means of (2), it can be shown that

$$\mathbf{e}^{(n)} = (\mathbf{I}-\omega\mathbf{L})^{-1}\,[\omega\mathbf{U}-(\omega-1)\,\mathbf{I}]\,\mathbf{e}^{(n-1)} = \mathbf{E}(\omega)\,\mathbf{e}^{(n-1)}, \text{ say.} \tag{24}$$

The essence of the method is to use an optimum value ω_b of ω which minimizes the modulus of the largest latent root of the iteration matrix $\mathbf{E}(\omega)$.

For suitable rearrangements of matrices possessing Property A, it can be shown [**112**] that

$$\omega_b = \frac{2}{1+\sqrt{(1-\theta^2)}}, \tag{25}$$

where θ is the largest latent root of $\mathbf{L}+\mathbf{U}$. Further, the largest latent root of $\mathbf{E}(\omega_b)$ has modulus ω_b-1, and is associated with a principal vector of grade 2.

In the example (3), $\theta = \frac{1}{2}$ and $\omega_b = 1{\cdot}0718$. The early iterations and displacements starting from $\mathbf{x}^{(0)} = 0$ are given below, and it can be verified that five-decimal accuracy is obtained with six cycles.

$\boldsymbol{\gamma}^{(0)}$	$\mathbf{x}^{(1)}$	$\boldsymbol{\gamma}^{(1)}$	$\mathbf{x}^{(2)}$	$\boldsymbol{\gamma}^{(2)}$	$\mathbf{x}^{(3)}$	$\boldsymbol{\gamma}^{(3)}$	$\mathbf{x}^{(4)}$
0·5	0·53590	0·23686	0·78977	0·07248	0·86745	0·00625	0·87415
0·63398	0·67950	0·15802	0·84887	0·02199	0·87244	0·00216	0·87476
0·38398	0·41155	0·17596	0·60014	0·02072	0·62235	0·00225	0·62476
0·52276	0·56029	0·05196	0·61598	0·00772	0·62425	0·00063	0·62493

For many large matrices θ is close to unity, and the superiority of successive overrelaxation is demonstrable as follows. Let θ^2 equal $1-\epsilon^2$, where ϵ is small. Then

$$\omega_b - 1 = \frac{1-\epsilon}{1+\epsilon} \fallingdotseq 1-2\epsilon. \tag{26}$$

When $\epsilon = 0{\cdot}1$, $\omega_b-1 = 0{\cdot}82$, while $\theta^2 = 0{\cdot}99$, and $\theta = 0{\cdot}995$.

11. One feature of these iterative methods is that the matrix of the original equations is used in unmodified form during the computation. This is particularly advantageous when the matrix contains many zero elements distributed systematically, as in finite-difference equations. Further, the non-zero elements in these matrices are often simple binary fractions, and this benefits both the programmer of an automatic computer with binary shift facilities, and the desk-machine worker. Moreover, it is convenient for the latter to work with a small number of figures in the early iterations, and to add figures as the approximation converges.

In addition to the classes of matrix cited in § 9, convergence is likely to be rapid when the off-diagonal elements are small compared with the diagonal elements. Even when such favourable conditions do not obtain, however, it is sometimes possible to accelerate convergence by applying Aitken's technique (see Chapter 13, § 3) to the corresponding elements of three successive vectors in a sequence of iterates. However, little benefit will result until the error consists mainly of the vector corresponding to the largest latent root of the iteration matrix, and it is not easy to provide an algorithm to determine when this stage has been reached.

Successive overrelaxation is very suitable for use on automatic computers. In particular, storage need be allocated for only one iteration vector, since each element of a new iterate may overwrite the corresponding element of the preceding approximation. It is not easy to determine the optimum overrelaxation parameter by an automatic sequence of arithmetical operations, but considerable progress has been made in the case of Property A matrices [**113**]. In addition, qualitative arguments often provide an estimate of ω_b which gives very good results [**114**].

5

LINEAR EQUATIONS AND MATRICES: ERROR ANALYSIS

INTRODUCTION

1. The problem of error analysis associated with the solution of n simultaneous linear equations

$$\mathbf{Ax} = \mathbf{b}, \tag{1}$$

may be regarded as the determination of the effect upon the solution $\mathbf{x}$ of (i) errors (if any) in the data, that is, the elements of $\mathbf{A}$ and $\mathbf{b}$, (ii) rounding errors introduced during the course of computing the solution. In the application of direct methods the number of multiplications is very large, approximately $\frac{1}{3}n^3$ for the methods of Chapter 1, and the number of roundings correspondingly great. The problem is therefore not easy; at the same time it is important.

The direct determination of the effect upon the solution of each individual rounding error is possible, but the total effort entailed greatly exceeds that required to compute the solutions themselves. Theoretical upper bounds for the errors obtained by this approach are not easy to calculate and also tend to overestimate grossly the actual errors.

2. We proceed here on different lines. After obtaining the approximate solution of equations (1), we seek perturbations (or rather upper bounds for perturbations) of the elements of $\mathbf{A}$ and $\mathbf{b}$ such that the computed solution is the *exact* solution of the perturbed equations.

We shall find that in this form the problem is quite tractable. From the results obtained we are able to make searching comparisons between the accuracy of the various methods of solution, and also arrive at practical rules governing the safe but economical number of guarding figures which need to be carried during the computation of the solution.

'ACCURATE' SOLUTIONS

3. Before proceeding with the error analysis proper, we ask the question: 'What is it reasonable to expect of our solution?'

Suppose that the elements of $\mathbf{A}$ and $\mathbf{b}$ do not exceed unity in absolute value (this can always be arranged by scaling), and that we use a working precision of t figures. Then, unless the elements of $\mathbf{A}$ and $\mathbf{b}$ are exact t-decimal numbers, they will have to be rounded to t decimals and we inevitably solve a perturbed set of equations

$$(\mathbf{A}+\delta\mathbf{A})\,\mathbf{x} = \mathbf{b}+\delta\mathbf{b}, \tag{2}$$

in which the upper bound of the elements of $\delta\mathbf{A}, \delta\mathbf{b}$ is $\frac{1}{2}10^{-t}$.

4. Even if the elements are exact t-decimal numbers, in the process of solution they will almost inevitably be multiplied by numbers which are not small integers. Consider the effect of the single simple operation of multiplying the equations (1) throughout by k, a t-digit number in the range 0·1 to 1·0. As a result of rounding to t decimals (1) is replaced by

$$(k\mathbf{A}+\mathbf{E})\mathbf{x} = k\mathbf{b}+\mathbf{e}, \tag{3}$$

where the elements of $\mathbf{E}$ and $\mathbf{e}$ are bounded by $\frac{1}{2}10^{-t}$. This is equivalent to

$$(\mathbf{A}+k^{-1}\mathbf{E})\mathbf{x} = \mathbf{b}+k^{-1}\mathbf{e}, \tag{4}$$

and the error terms of this perturbed set have bounds which depend on k but are at least $\frac{1}{2}10^{-t}$.

5. We shall show that the approximate solution of a set of equations (1) obtained by a direct method is the exact solution of a perturbed set of the form (2). It is clear from the simple illustrations that have just been given that it would be quite unreasonable to expect the bounds of the elements of $\delta\mathbf{A}$ and $\delta\mathbf{b}$ to be less than $\frac{1}{2}10^{-t}$, for a working precision of t figures. Much higher bounds might well be anticipated and it is a matter for some surprise that with certain methods the ideal is in fact approached.

ERROR ANALYSIS FOR GAUSSIAN ELIMINATION

6. We introduce our method of error analysis in terms of the numerical example of § 8 of Chapter 1. Purely for notational convenience we have rearranged the original equations so that successive pivotal rows are in order one beneath the other. This use of hindsight does not affect the analysis; it merely eases the description.

On the left of Table 1 we repeat the numbers recorded during the computation of the solution by Gaussian elimination; for convenience we have included previous pivotal rows in each reduced set and also elements that have become zero. We therefore have four sets of equations

$$\mathbf{A}^{(r)}\mathbf{x} = \mathbf{b}^{(r)} \quad (r = 1, 2, 3, 4), \tag{5}$$

all of which would have had identical solutions if the computation had been performed exactly.

Reduction to triangular form

7. We consider first the reduction to triangular form. We work backwards from the fourth set of equations deriving successively perturbations to the third, second and first sets so that they are satisfied exactly by the exact solutions of the fourth (triangular) set. Such perturbations are not unique and we select those which reproduce exactly the recorded multipliers m_{ij}. The perturbed sets of equations are denoted by

$$(\mathbf{A}^{(r)}+\delta\mathbf{A}^{(r)})\mathbf{x} = \mathbf{b}^{(r)}+\delta\mathbf{b}^{(r)} \quad (r = 1, 2, 3). \tag{6}$$

TABLE 1

	$A^{(1)}$				$b^{(1)}$	$10^5\,\delta A^{(1)}$				$10^5\,\delta b^{(1)}$
	0·4096	*0·1234*	*0·3678*	*0·2943*	*0·4043*	0	0	0	0	0
$m_{21} = -0{\cdot}43555$	0·1784	0·4002	0·2786	0·3927	−0·2557	0·12800	−0·31300	−0·47100	0·23650	0·28650
$m_{31} = -0{\cdot}54834$	0·2246	0·3872	0·4015	0·1129	0·1550	0·00640	−0·40090	−0·02280	−0·67420	0·39450
$m_{41} = -0{\cdot}88989$	0·3645	0·1920	0·3728	0·0643	0·4240	−0·10560	0·07595	−0·05116	1·12126	−0·36287

	$A^{(2)}$				$b^{(2)}$	$10^5\,\delta A^{(2)}$				$10^5\,\delta b^{(2)}$
	0·4096	0·1234	0·3678	0·2943	0·4043	0	0	0	0	0
	0·00000	*0·34645*	*0·11840*	*0·26452*	*−0·43179*	0	0	0	0	0
$m_{32} = -0{\cdot}92230$	0·00000	0·31953	0·19982	−0·04848	−0·06669	0	0·08350	0·03200	−0·32040	0·00830
$m_{42} = -0{\cdot}23723$	0·00000	0·08219	0·04550	−0·19759	0·06422	0	−0·16665	−0·20536	0·65856	−0·61557

	$A^{(3)}$				$b^{(3)}$	$10^5\,\delta A^{(3)}$				$10^5\,\delta b^{(3)}$
	0·4096	0·1234	0·3678	0·2943	0·4043	0	0	0	0	0
	0·00000	0·34645	0·11840	0·26452	−0·43179	0	0	0	0	0
	0·00000	*0·00000*	*0·09062*	*−0·29245*	*0·33155*	0	0	0	0	0
$m_{43} = -0{\cdot}19212$	0·00000	0·00000	0·01741	−0·26034	0·16665	0	0	−0·00856	0·45060	−0·26140

$A^{(4)}$				$b^{(4)}$
0·4096	0·1234	0·3678	0·2943	0·4043
0·00000	0·34645	0·11840	0·26452	−0·43179
0·00000	0·00000	0·09062	−0·29245	0·33155
0·00000	*0·00000*	*0·00000*	*−0·20415*	*0·10295*

If (x_1, x_2, x_3, x_4) is the *exact* solution of the *computed* triangular set $\mathbf{A}^{(4)}\mathbf{x} = \mathbf{b}^{(4)}$, we have

$$a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 \equiv b_1^{(1)}, \tag{7}$$

$$a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + a_{24}^{(2)}x_4 \equiv b_2^{(2)}, \tag{8}$$

$$a_{33}^{(3)}x_3 + a_{34}^{(3)}x_4 \equiv b_3^{(3)}, \tag{9}$$

$$a_{44}^{(4)}x_4 \equiv b_4^{(4)}, \tag{10}$$

where equivalence signs denote that the equations are satisfied exactly. As is indicated by the superscripts, the pivotal row (7) of the first set of equations appears unaltered in sets 2, 3, 4, (8) occurs in sets 2, 3, 4 and (9) in 3 and 4.

In the subsequent analysis we shall assume that none of the elements $a_{ij}^{(r)}$ and $b_i^{(r)}$ exceeds unity. This is clearly true in the example, and is a normal feature of most automatic work. It is easily arranged, because any growth of the maximum coefficient from set to set is quite slow in practice and a preliminary scaling which reduces all the original coefficients $a_{ij}^{(1)}$ to a maximum of, say, $\frac{1}{8}$ is sufficient in all but pathological cases. This point is also discussed in Chapter 2, § 6.

The determination of $\delta\mathbf{A}^{(3)}$ *and* $\delta\mathbf{b}^{(3)}$

8. Equations (7), (8) and (9) are the same in set 3 as in set 4 and so are automatically satisfied exactly. The multiplier m_{43} and elements $a_{44}^{(4)}$ and $b_4^{(4)}$ were obtained from the equations

$$m_{43} = -a_{43}^{(3)}/a_{33}^{(3)}, \tag{11}$$

$$a_{44}^{(4)} = a_{44}^{(3)} + m_{43}a_{34}^{(3)}, \tag{12}$$

$$b_4^{(4)} = b_4^{(3)} + m_{43}b_3^{(3)}, \tag{13}$$

and rounded to five decimals. The rounded m_{43} accordingly satisfies exactly

$$-m_{43} \equiv (a_{43}^{(3)}/a_{33}^{(3)}) + \eta_{43}, \qquad |\eta_{43}| \leqslant \tfrac{1}{2}10^{-5}, \tag{14}$$

whence

$$-m_{43}a_{33}^{(3)} \equiv a_{43}^{(3)} + \epsilon_{43}^{(3)}, \tag{15}$$

where

$$|\epsilon_{43}^{(3)}| = |a_{33}^{(3)}\eta_{43}| \leqslant |\eta_{43}| \leqslant \tfrac{1}{2}10^{-5}. \tag{16}$$

The values of $a_{44}^{(4)}$ and $b_4^{(4)}$ obtained from equations (12) and (13) are rounded to five decimals, so that

$$a_{44}^{(4)} \equiv a_{44}^{(3)} + m_{43}a_{34}^{(3)} + \epsilon_{44}^{(3)}, \qquad |\epsilon_{44}^{(3)}| \leqslant \tfrac{1}{2}10^{-5}, \tag{17}$$

$$b_4^{(4)} \equiv b_4^{(3)} + m_{43}b_3^{(3)} + \epsilon_4^{(3)}, \qquad |\epsilon_4^{(3)}| \leqslant \tfrac{1}{2}10^{-5}. \tag{18}$$

$\epsilon_{43}^{(3)}$, $\epsilon_{44}^{(3)}$ and $\epsilon_4^{(3)}$ are thus the perturbations required in $a_{43}^{(3)}$, $a_{44}^{(3)}$ and $b_4^{(3)}$ to make the third set exactly equivalent to the fourth and to reproduce exactly the computed multiplier m_{43}. Each perturbation is bounded by $\frac{1}{2}10^{-5}$; this is obviously true for $\epsilon_{44}^{(3)}$ and $\epsilon_4^{(3)}$, and $\epsilon_{43}^{(3)}$ is a rounding error multiplied by a number which, by hypothesis, does not exceed unity. Thus

$$[\delta\mathbf{A}^{(3)} \,|\, \delta\mathbf{b}^{(3)}] \leqslant \tfrac{1}{2}10^{-5}\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{array}\right], \tag{19}$$

where the inequality sign means that the absolute value of each element of the left-hand side does not exceed the corresponding element of the right-hand side. In the example of Table 1 the computations corresponding to (15), (17) and (18) are

$$\begin{aligned}
\delta a_{43}^{(3)} = \epsilon_{43}^{(3)} &= -(-0{\cdot}19212)(0{\cdot}09062) - 0{\cdot}01741 \\
&= -10^{-5}(0{\cdot}00856), \\
\delta a_{44}^{(3)} = \epsilon_{44}^{(3)} &= -0{\cdot}20415 + 0{\cdot}26034 - (-0{\cdot}19212)(-0{\cdot}29245) \\
&= 10^{-5}(0{\cdot}45060), \\
\delta b_{4}^{(3)} = \epsilon_{4}^{(3)} &= 0{\cdot}10295 - 0{\cdot}16665 - (-0{\cdot}19212)(0{\cdot}33155) \\
&= -10^{-5}(0{\cdot}26140).
\end{aligned}$$

It may be noted that these numbers are exact ten-decimal numbers.

The determination of $\delta\mathbf{A}^{(2)}$ *and* $\delta\mathbf{b}^{(2)}$

9. Equations (7) and (8) appear unchanged in set 2 and so are automatically satisfied. Since the third equation of the third set is not perturbed, the perturbations of the third equation of the second set are obtained by an analysis following precisely the lines of the previous section. For example, $\epsilon_3^{(2)}$ is obtained from the equation

$$b_3^{(3)} \equiv b_3^{(2)} + m_{32} b_2^{(2)} + \epsilon_3^{(2)}, \qquad |\epsilon_3^{(2)}| \leqslant \tfrac{1}{2}10^{-5}, \tag{20}$$

which corresponds exactly to (18). Therefore $\epsilon_{32}^{(2)}$, $\epsilon_{33}^{(2)}$, $\epsilon_{34}^{(2)}$ and $\epsilon_3^{(2)}$ are all bounded by $\frac{1}{2}10^{-5}$.

The perturbation of the fourth equation of set 2 must, however, be chosen so that the *perturbed* fourth equation of set 3 is exactly reproduced. Thus while $b_4^{(3)}$ satisfies the relation

$$b_4^{(3)} \equiv b_4^{(2)} + m_{42} b_2^{(2)} + \epsilon_4^{(2)}, \qquad |\epsilon_4^{(2)}| \leqslant \tfrac{1}{2}10^{-5}, \tag{21}$$

it is $b_4^{(3)} + \epsilon_4^{(3)}$ that must be reproduced. Since

$$b_4^{(3)} + \epsilon_4^{(3)} \equiv b_4^{(2)} + m_{42} b_2^{(2)} + \epsilon_4^{(2)} + \epsilon_4^{(3)}, \tag{22}$$

we see that $\epsilon_4^{(2)} + \epsilon_4^{(3)}$ is the perturbation that must be added to $b_4^{(2)}$; thus the perturbation arises from two rounding errors which are purely additive and do not interact in any way. This is also true for the perturbations to $a_{43}^{(2)}$ and $a_{44}^{(2)}$. Since the zero element $a_{42}^{(3)}$ was unperturbed, we see that the necessary perturbation of $a_{42}^{(2)}$ arises from a single rounding error and is given by the equation

$$-m_{42} a_{22}^{(2)} = a_{42}^{(2)} + \epsilon_{42}^{(2)}, \tag{23}$$

exactly analogous to (15). Thus

$$[\delta\mathbf{A}^{(2)} \,|\, \delta\mathbf{b}^{(2)}] \leqslant \tfrac{1}{2}10^{-5} \left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 \end{array}\right]. \tag{24}$$

The right-hand side of (24) may be regarded as the sum of the bounds of the perturbation (19) arising in the reduction from the third to the fourth set, and the corresponding perturbations, bounded by

$$\tfrac{1}{2}10^{-5}\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{array}\right], \tag{25}$$

arising in the reduction from the second to the third set, the two sets of perturbations being *simply additive without interaction.*

In the example of Table 1 the computation corresponding to (22), for example, is

$$\begin{aligned} \delta b_4^{(2)} &= \epsilon_4^{(2)} + \epsilon_4^{(3)} \\ &= 0{\cdot}16665 - 0{\cdot}06422 + (-0{\cdot}23723)(0{\cdot}43179) - 10^{-5}(0{\cdot}26140) \\ &= -10^{-5}(0{\cdot}61557). \end{aligned}$$

Similar calculations lead to the perturbations given in Table 1. Again they are all terminating numbers of ten decimal places.

The determination of $\delta\mathbf{A}^{(1)}$ *and* $\delta\mathbf{b}^{(1)}$

10. The pattern of the analysis is now quite clear. The next step leads to an additional set of perturbations bounded by

$$\tfrac{1}{2}10^{-5}\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{array}\right] \tag{26}$$

to be added to (24), so that

$$[\delta\mathbf{A}^{(1)}\,|\,\delta\mathbf{b}^{(1)}] \leqslant \tfrac{1}{2}10^{-5}\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \end{array}\right]. \tag{27}$$

The results for our example are again given in Table 1; it will be seen that all the perturbations lie within the limits (27).

The back-substitution

11. We have shown that a perturbed set of equations

$$(\mathbf{A}^{(1)} + \delta\mathbf{A}^{(1)})\mathbf{x} = \mathbf{b}^{(1)} + \delta\mathbf{b}^{(1)}, \tag{28}$$

can be obtained such that successive derived sets are exactly

$$(\mathbf{A}^{(2)} + \delta\mathbf{A}^{(2)})\mathbf{x} = \mathbf{b}^{(2)} + \delta\mathbf{b}^{(2)}, \tag{29}$$

$$(\mathbf{A}^{(3)} + \delta\mathbf{A}^{(3)})\mathbf{x} = \mathbf{b}^{(3)} + \delta\mathbf{b}^{(3)}, \tag{30}$$

$$\mathbf{A}^{(4)}\mathbf{x} = \mathbf{b}^{(4)}, \tag{31}$$

and the multipliers used in the reduction are exactly the computed multipliers; the actual perturbations for our example are given in Table 1 and in the general case of order 4 they have bounds given by (27), (24) and (19).

We now consider the solution of the triangular set (31), written in full in equations (7) to (10). Consider, for example, the determination of x_2, when x_3 and x_4 have been found. Equation (8) is used in the form

$$x_2 \equiv \frac{b_2^{(2)} - a_{23}^{(2)} x_3 - a_{24}^{(2)} x_4}{a_{22}^{(2)}} + \eta_2, \qquad |\eta_2| \leqslant \tfrac{1}{2}10^{-5}, \tag{32}$$

in which there is only one rounding error, η_2, since the numerator is accumulated exactly on the machine.

In taking $|\eta_2| \leqslant \frac{1}{2}10^{-5}$ we are assuming that the x_i are computed and recorded to 5 *decimals*. Since the x_i may be large this could mean working with more than 5 significant figures; we return to this point in § 15. From (32) we obtain

$$a_{22}^{(2)} x_2 + a_{23}^{(2)} x_3 + a_{24}^{(2)} x_4 \equiv b_2^{(2)} + \delta c_2, \tag{33}$$

where

$$|\delta c_2| = |\eta_2 a_{22}^{(2)}| \leqslant |\eta_2| \leqslant \tfrac{1}{2}10^{-5}. \tag{34}$$

A similar result holds for each variable. The computed solution of the triangular set of equations is therefore the exact solution of the equations

$$\mathbf{A}^{(4)}\mathbf{x} = \mathbf{b}^{(4)} + \delta\mathbf{c}, \qquad |\delta c_i| \leqslant \tfrac{1}{2}10^{-5}. \tag{35}$$

Now it is readily seen that the addition of a further perturbation $\delta\mathbf{b}$ given by

$$\delta\mathbf{b} = \begin{bmatrix} \delta c_1 \\ -m_{21}\delta c_1 + \delta c_2 \\ -m_{31}\delta c_1 - m_{32}\delta c_2 + \delta c_3 \\ -m_{41}\delta c_1 - m_{42}\delta c_2 - m_{43}\delta c_3 + \delta c_4 \end{bmatrix} \tag{36}$$

to the right-hand side of (28) will result in the final reduced set of equations having the form (35). This follows because the perturbation $\delta\mathbf{b}$ does not affect the multipliers, and the consequent additions to the right-hand sides of (29), (30) and (31) are respectively

$$\begin{bmatrix} \delta c_1 \\ \delta c_2 \\ -m_{32}\delta c_2 + \delta c_3 \\ -m_{42}\delta c_2 - m_{43}\delta c_3 + \delta c_4 \end{bmatrix}, \quad \begin{bmatrix} \delta c_1 \\ \delta c_2 \\ \delta c_3 \\ -m_{43}\delta c_3 + \delta c_4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \delta c_1 \\ \delta c_2 \\ \delta c_3 \\ \delta c_4 \end{bmatrix}.$$

Since no multiplier exceeds unity,

$$\delta\mathbf{b} \leqslant \tfrac{1}{2}10^{-5} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}. \tag{37}$$

In the example of Table 1, we find

$$10^5\,\delta b_1 = -0{\cdot}00790, \qquad 10^5\,\delta b_2 = 0{\cdot}02252\ 91550,$$
$$10^5\,\delta b_3 = -0{\cdot}01306\ 97550, \qquad 10^5\,\delta b_4 = 0{\cdot}07320\ 03293.$$

Summarizing, we have shown that the *computed* solution is the *exact* solution of the set of equations

$$(\mathbf{A}^{(1)} + \delta\mathbf{A}^{(1)})\mathbf{x} = \mathbf{b}^{(1)} + \delta\mathbf{b}^{(1)} + \delta\mathbf{b}, \tag{38}$$

where

$$[\delta\mathbf{A}^{(1)}\,|\,\delta\mathbf{b}^{(1)}\,|\,\delta\mathbf{b}] \leqslant \tfrac{1}{2}10^{-5}\left[\begin{array}{cccc|c|c} 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 2 & 2 & 3 \\ 1 & 2 & 3 & 3 & 3 & 4 \end{array}\right], \tag{39}$$

and we have kept separate the perturbations to the right-hand side arising from the reduction and the back-substitution.

The general case

12. The result (39) can be extended immediately to a set of n equations, solved using a working precision of t decimals. The bounds of the perturbations are given by

$$[\delta\mathbf{A}^{(1)}\,|\,\delta\mathbf{b}^{(1)}\,|\,\delta\mathbf{b}] \leqslant \tfrac{1}{2}10^{-t}\left[\begin{array}{cccccccc|c|c} 0 & 0 & 0 & 0 & \ldots & 0 & 0 & & 0 & 1 \\ 1 & 1 & 1 & 1 & \ldots & 1 & 1 & & 1 & 2 \\ 1 & 2 & 2 & 2 & \ldots & 2 & 2 & & 2 & 3 \\ . & . & . & . & \ldots & . & . & & . & . \\ . & . & . & . & \ldots & . & . & & . & . \\ 1 & 2 & 3 & 4 & \ldots & (n-2) & (n-2) & & n-2 & n-1 \\ 1 & 2 & 3 & 4 & \ldots & (n-1) & (n-1) & & n-1 & n \end{array}\right]. \tag{40}$$

When binary arithmetic is used, 10^{-t} is replaced by 2^{-t}.

ERROR ANALYSIS FOR TRIANGULAR DECOMPOSITION

13. A similar analysis can be applied to the method of triangular decomposition with interchanges, described in Chapter 2, §§ 15, 16. If t decimals are retained throughout, the computed solution $\mathbf{x}$ satisfies exactly a set of equations of the form (38), with

$$|\delta a_{ij}^{(1)}| \leqslant \tfrac{1}{2}10^{-t}, \qquad |\delta b_i^{(1)}| \leqslant \tfrac{1}{2}10^{-t}, \qquad |\delta b_i| \leqslant \tfrac{1}{2}i10^{-t}. \tag{41}$$

The proof [32] is appreciably simpler than that for the method of Gaussian elimination, and the smaller bounds obtained are very impressive. In fact, as foreshadowed in § 5, we have here a method which approaches very closely what must be considered the ideal in which the bound for each perturbation is $\tfrac{1}{2}10^{-t}$.

Again, when binary arithmetic is used, 10^{-t} is replaced by 2^{-t}.

14. Let us suppose that the elements of $\mathbf{A}, \mathbf{b}$ in the original equations are prescribed to t decimals. Then, since the largest bound of the perturbations given by (40) is the sum of n rounding errors, the retention of one extra decimal for $n = 10$, two for $n = 100$ and in general $\log_{10} n$, will result in the solution obtained by Gaussian elimination being the exact solution of a set of equations which do not differ from the original set by more than the possible rounding errors inherent in the data. For binary arithmetic, $\log_2 n$ additional binary places will be required to ensure this.

This result is also true for the method of triangular decomposition. However, since in this case the elements of $\delta\mathbf{A}^{(1)}, \delta\mathbf{b}^{(1)}$ are bounded by one rounding error, it is necessary to carry the additional decimals only in the back-substitution, which is much the smaller part of the calculation.

It should be borne in mind that (40) and (41) give strict upper bounds. In practice we expect the rounding errors to accumulate statistically and, consequently, n to be replaced by $\sqrt{n}$. This suggests that it would not be unreasonable to use only one additional guarding figure for sets of up to 100 equations.

15. We return now to the point made in § 11 regarding the size of the solution $\mathbf{x}$. If we are restricted by our machine to the use of t significant figures, and if the largest element x_t lies between 10^k and 10^{k-1}, then we are only able to retain $t-k$ decimals in the course of the back-substitution. This has the effect of multiplying the associated perturbation $\delta\mathbf{b}$ by 10^k, so that for both Gaussian elimination and triangular decomposition we now have the bounds

$$|\delta b_i| \leqslant \tfrac{1}{2} i 10^{k-t}. \tag{42}$$

The consequent loss of accuracy is not as great as might appear. It can be shown, for example, that it is possible to absorb $\delta\mathbf{b}$ by an additional perturbation of $\mathbf{A}$ which does not contain the factor 10^k. Alternatively, we may consider the upper bounds for the residuals.

The residual vector $\mathbf{r}$ is given by

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x} = -\delta\mathbf{b}^{(1)} - \delta\mathbf{b} + \delta\mathbf{A}^{(1)}\mathbf{x}. \tag{43}$$

In the case of Gaussian elimination it follows from (40), and the fact that each $|x_i|$ is bounded by 10^k, that

$$\begin{aligned} |r_i| &\leqslant \tfrac{1}{2}(i-1)\,10^{-t} + \tfrac{1}{2} i 10^{-t} + \tfrac{1}{2} 10^{k-t}[1+2+\ldots+(i-1)+(n-i+1)\,(i-1)] \\ &= \tfrac{1}{2}(i-1)\,10^{-t} + \tfrac{1}{2} i 10^{-t} + \tfrac{1}{2} 10^{k-t}(i-1)\,(n-\tfrac{1}{2}i+1) \end{aligned} \tag{44}$$

when we retain t decimals in the back-substitution, and

$$|r_i'| \leqslant \tfrac{1}{2}(i-1)\,10^{-t} + \tfrac{1}{2} i 10^{k-t} + \tfrac{1}{2} 10^{k-t}(i-1)\,(n-\tfrac{1}{2}i+1) \tag{45}$$

when we retain $t-k$ decimals in the back-substitution. The difference between these bounds is less than $\tfrac{1}{2} n 10^{k-t}$ while each is greater than $\tfrac{1}{4} n^2\, 10^{k-t}$ when $i = n$, a proportional increase of less than $2/n$.

In the case of triangular decomposition the bounds corresponding to (44) and (45) are

$$|r_i| \leqslant \tfrac{1}{2} 10^{-t} + \tfrac{1}{2} i 10^{-t} + \tfrac{1}{2} n 10^{k-t} \tag{46}$$

and $$|r_i'| \leqslant \tfrac{1}{2}10^{-t} + \tfrac{1}{2}i10^{k-t} + \tfrac{1}{2}n10^{k-t}; \tag{47}$$

in this case the maximum bound for $|r_i'|$ is less than twice that for $|r_i|$.

16. These bounds may be compared with those obtained for the residuals when we take the exact solution $\mathbf{x}_e$, round it to $k-t$ decimals, and substitute this rounded solution $\mathbf{x}_r$ into the original equations. We then have

$$|(\mathbf{A}\mathbf{x}_r - \mathbf{b})_i| = |[\mathbf{A}(\mathbf{x}_r - \mathbf{x}_e)]_i| \leqslant \tfrac{1}{2}n10^{k-t}. \tag{48}$$

We may replace (44) and (45) by the slightly weaker inequality

$$|r_i| \leqslant \tfrac{1}{2}10^{k-t}[(i-1)(n-\tfrac{1}{2}i+3)+1].$$

The largest bound is that for $|r_n|$. Hence, for all i,

$$|r_i| \leqslant \tfrac{1}{2}10^{k-t}(\tfrac{1}{2}n^2 + \tfrac{5}{2}n - 2) \quad \text{(Gaussian elimination).} \tag{49}$$

Similarly, for all i,

$$|r_i| \leqslant \tfrac{1}{2}10^{k-t}(2n+1) \quad \text{(triangular decomposition).} \tag{50}$$

The ratio of the bounds (49), (48) is

$$(n^2 + 5n - 4)/2n, \tag{51}$$

which is an increasing function of n and attains the values 10, 100 for $n \doteqdot 15$, 195 respectively. Thus for sets of order up to 15 the residuals corresponding to the solution computed by Gaussian elimination may be expected to be no greater than those corresponding to the exact solution rounded to one decimal less, while for sets of orders 15 to 195, no greater than those corresponding to the exact solution rounded to two decimals less.

For triangular resolution we have the remarkable result that for any order the residuals are not expected to exceed twice those of the exact solution rounded to the same number of decimals.

17. Again the results obtained in the preceding sections are strict upper bounds. In practice we would expect a statistical accumulation of roundings, in which case the ratio (51) would be replaced approximately by its square root. In this case we could expect the residuals to be no greater than those of the exact solution rounded to one decimal less for sets of orders up to 195.

These results are even more satisfactory than those given in § 14. The difference lies in the fact that in § 14 we attached as much weight to the perturbations $\delta\mathbf{b}$ as to $\delta\mathbf{A}$. An assessment based purely on the size of the residuals is much more realistic. If $\mathbf{r}$ is the residual vector corresponding to an approximate solution $\mathbf{x}$, then the true solution is $\mathbf{x} + \mathbf{A}^{-1}\mathbf{r}$; the error is therefore directly dependent on the residual vector.

THE EFFECT OF USING LARGE MULTIPLIERS

18. We have stressed the importance of using as pivots the largest element in each column, that is, carrying out our calculations with interchanges. It may be thought that this restriction is really unnecessary and could be overcome, for example, by using floating-point arithmetic. The example, given in Table 2 opposite, shows that this is not so.

The original equations and first reduced set are denoted by

$$\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}, \qquad \mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}, \tag{52}$$

respectively, as in the previous example. The smallest element, a_{11}, is used as pivot in the reduction. The 'perturbed' form of the original equations which corresponds exactly to the reduced set is given by the third array of numbers, computed by the method of § 8. The second and third members differ substantially from the original ones; indeed, their solution (which is, of course, also the solution of the reduced set) bears no resemblance to that of the original equations.

TABLE 2

m	$\mathbf{A}^{(1)}$			$\mathbf{b}^{(1)}$
	0·000003	*0·213472*	*0·332147*	*0·235262*
−71837·3	0·215512	0·375623	0·476625	0·127653
−57752·3	0·173257	0·663257	0·625675	0·285321
	$\mathbf{A}^{(2)}$			$\mathbf{b}^{(2)}$
	0·000003	0·213472	0·332147	0·235262
	0·000000	−15334·9	−23860·1	−16900·5
	0·000000	−12327·8	−19181·6	−13586·6
	Exact equivalent of first reduced set			
	0·000003	0·213472	0·332147	0·235262
	0·2155129	0·3521056	0·4436831	0·0868726
	0·1732569	0·6989856	0·6531881	0·3216026

What has happened is that the second and third equations of the reduced set are almost entirely composed of a multiple of the pivotal equation. We would obtain an identical system $\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$ for a whole class of matrices $[\mathbf{A}^{(1)} \mid \mathbf{b}^{(1)}]$. For consider the computation of a typical element $a_{23}^{(2)}$ in six-figure floating-point arithmetic, given by

$$\begin{aligned} a_{23}^{(2)} &= a_{23}^{(1)} + m_{21} a_{13}^{(1)} \\ &= 0{\cdot}476625 - 71837{\cdot}3 \times 0{\cdot}332147 \\ &= 0{\cdot}476625 - 23860{\cdot}6 = -23860{\cdot}1. \end{aligned}$$

Because the product $m_{21} a_{13}^{(1)}$ is so much larger than $a_{23}^{(1)}$, nearly all figures of the latter are ignored. We obtain precisely the same $a_{23}^{(2)}$ for all values of $a_{23}^{(1)}$ lying between 0·45 and 0·55.

ILL-CONDITIONED EQUATIONS

19. It may well be asked what bearing the results obtained in this chapter have on the difficulties associated with ill-conditioned equations. We have obtained bounds for the perturbations in $\mathbf{A}$ and $\mathbf{b}$ which give equations satisfied exactly by the computed $\mathbf{x}$, and we have been able to deduce rules for the number of guarding decimals that should, if possible, be used.

If, as is usually the case, the coefficients in the original equations are approximate, then we can determine immediately from these bounds whether or not the computed solution is *as accurate as the data warrants.* This is the case for both ill- and well-conditioned equations and in many problems this information will suffice.

20. In some applications, however, we may wish to know to how many figures our solutions actually agree with the exact solution of the original equations. An analytical investigation of this requires a knowledge of the inverse matrix $\mathbf{A}^{-1}$, and even the estimation of a satisfactory upper bound for its elements will usually involve more effort than the solution of the equations. In practice it will be necessary to follow the procedure described in Chapter 2, § 18.

ERROR ANALYSIS OF OTHER ALGEBRAIC PROCESSES

21. The technique described in this chapter has been applied to a wide range of algebraic processes, including many of the better-known methods for solving linear equations, computing eigensystems and calculating the zeros of polynomials [**216**]. Both fixed-point and floating-point computation have been considered. In all cases the computed solution of the problem has been shown to be the exact solution of a perturbed problem, and bounds have been determined for the perturbations. This has the advantage of allowing a direct comparison between the effect of rounding errors and of the errors inherent in the original data.

If small, strict bounds can be found for the perturbations, then the method being analysed must be regarded as a good method. If such bounds cannot be found, however, we cannot necessarily deduce that the method is bad. Nevertheless, in such a case the analysis will frequently suggest cases in which the method will undoubtedly lead to inaccurate results. An example of this kind has been given in § 18, where we showed the importance of interchanges in Gaussian elimination.

Sometimes the bounds for the perturbations will provide strict *a priori* bounds for the computed solution; the most important example of this is the calculation of the latent roots of symmetric matrices. More frequently, however, the determination of a strict bound for the error in the solution will require information which is at least as difficult to obtain as the solution itself; in this case the best we can hope for is an *a posteriori* bound.

6

ZEROS OF POLYNOMIALS

EVALUATION OF A POLYNOMIAL

1. We take as our standard form of polynomial

$$f(z) = a_n z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \dots + a_0, \tag{1}$$

and suppose throughout this chapter that the coefficients a_s are real.

The evaluation of $f(z)$ for a given real value $z = \alpha$ can be effected on a desk machine equipped with a transfer, by cyclic repetition of the processes of multiplication, setting, addition and transference, according to the formula

$$f(\alpha) = [\{(a_n \alpha + a_{n-1}) \alpha + a_{n-2}\} \alpha + a_{n-3}] \alpha + \dots. \tag{2}$$

No intermediate recording is then necessary. This process is sometimes called *nested multiplication.*

2. The numbers which appear in the product register immediately before the successive transfers are the coefficients in the quotient polynomial which results on dividing $f(z)$ by the linear factor $z - \alpha$. For, by equating coefficients in the relation

$$f(z) = (q_n z^{n-1} + q_{n-1} z^{n-2} + \dots + q_1)(z - \alpha) + q_0, \tag{3}$$

we find

$$q_n = a_n, \qquad q_{n-1} = q_n \alpha + a_{n-1}, \tag{4}$$

and generally,

$$q_s = q_{s+1} \alpha + a_s \quad (s = n-1, n-2, \dots, 0). \tag{5}$$

Also

$$f(\alpha) = q_0. \tag{6}$$

POLYNOMIALS OF LOW DEGREE

3. Real roots of the equation

$$f(z) = 0, \tag{7}$$

that is, real *zeros* of $f(z)$, can be located by examining the sign of $f(z)$ at $z = 0$ and $\pm\infty$, and evaluating $f(z)$ at a few trial values of z. If $f(z)$ has opposite signs at $z = z_1$ and $z = z_2$, then at least one root lies between z_1 and z_2.

Regarding z_1 and z_2 as approximations to the root, we can obtain a new approximation z_3 by inverse linear interpolation, according to the formula

$$z_3 = \frac{z_1 f(z_2) - z_2 f(z_1)}{f(z_2) - f(z_1)} = z_2 - \frac{(z_2 - z_1) f(z_2)}{f(z_2) - f(z_1)}. \tag{8}$$

The process can then be repeated. Alternatively, Newton's rule, described in § 9 below, may be used.

4. If $f(z)$ is a cubic polynomial, one real root may be determined by this method and the corresponding linear factor then divided out, as described in § 2. The zeros of the residual quadratic, whether real or complex, may be obtained by application of the usual formula. This method is quite satisfactory for the solution of the occasional cubic equation.

5. Equations of higher degree may also be solved by this method, provided that at most one pair of roots is complex. If two pairs are complex, the problem reduces to that of solving a quartic equation. We now describe a variant of the classical method for solving equations of this degree which is convenient for computation.

THE QUARTIC

6. We suppose that $f(z)$ has been normalized so that the coefficient of z^4 is unity, and that

$$z^4+a_3z^3+a_2z^2+a_1z+a_0 = (z^2+b_1z+c_1)(z^2+b_2z+c_2). \qquad (9)$$

The problem is essentially to determine b_1, c_1, b_2, c_2 from a_0, a_1, a_2, a_3.

Equating coefficients, we obtain

$$a_3 = b_1+b_2, \qquad a_2 = b_1b_2+c_1+c_2, \qquad (10)$$

$$a_1 = b_1c_2+b_2c_1, \qquad a_0 = c_1c_2. \qquad (11)$$

We write

$$m = c_1+c_2. \qquad (12)$$

Then m satisfies the equation

$$m^3-a_2m^2+(a_1a_3-4a_0)m+(4a_2-a_3^2)a_0-a_1^2 = 0. \qquad (13)$$

For, by solving the first of (10) and the first of (11) for b_1 and b_2, we find

$$b_1 = \frac{a_1-a_3c_1}{c_2-c_1}, \qquad b_2 = \frac{a_3c_2-a_1}{c_2-c_1}. \qquad (14)$$

Substitution in the second of (10) gives

$$(c_1+c_2-a_2)(c_2-c_1)^2+a_1a_3(c_1+c_2)-a_3^2c_1c_2-a_1^2 = 0. \qquad (15)$$

Eliminating c_1 and c_2 from this equation by means of (12) and the second of (11), we immediately obtain (13).

Equation (13) is called the *reducing cubic*. A real root can be found by the method of § 3. Equation (12) and the second of (11) show that c_1 and c_2 are the roots of the quadratic equation

$$c^2-mc+a_0 = 0, \qquad (16)$$

and equations (14) then give b_1 and b_2.

In solving (13), a root should be sought which satisfies the inequality $m^2 \geqslant 4a_0$, otherwise the roots of (16) will be complex, and the factorization (9) will not be into conjugate pairs of roots. Thus if a_0 is positive, we first seek a value of m in the range $(2\sqrt{a_0}, \infty)$. If no such root exists, and this can only happen when some of the roots of the quartic equation are real, we seek a root of (13) in the range $(-\infty, -2\sqrt{a_0})$, which must exist.

Example

7. Consider the equation

$$z^4 - 2z^3 + 7z^2 - 10z + 11 = 0. \qquad (17)$$

Here $a_0 = 11, a_1 = -10, a_2 = 7, a_3 = -2$. Equation (13) becomes

$$\phi(m) \equiv m^3 - 7m^2 - 24m + 164 = 0. \qquad (18)$$

The value of $2\sqrt{a_0}$ is $2\sqrt{11} = 6{\cdot}63\ldots$. Taking the nearest integer value $m = 7$, we find $\phi(7) = -4$. This has the opposite sign to $\phi(+\infty)$; hence there is a root of (18) in the range $(7, \infty)$. Next, we find $\phi(8) = 36$; therefore the wanted root must be less than 8. Starting from these two approximations, we determine the root accurately to 5 decimals, say, by repeated application of (8):

m	7	8	7·1	7·151	7·14758	7·14766
$\phi(m)$	-4	36	$-1{\cdot}359$	0·09766	$-0{\cdot}00237$	$-0{\cdot}00003$

Equation (16) becomes

$$c^2 - 7{\cdot}14766c + 11 = 0, \qquad (19)$$

giving $c_1 = 2{\cdot}24257, c_2 = 4{\cdot}90509$. From (14) we obtain $b_1 = -2{\cdot}07129$, $b_2 = 0{\cdot}07129$. As a check, we verify that equations (10) and (11) are satisfied.

8. Polynomials having more than two pairs of complex zeros may be classified with the 'high-degree' polynomials. As a preliminary to outlining methods for finding the zeros of these polynomials we now describe the iterative processes of Newton and Bairstow.

NEWTON'S RULE

9. Let α be an approximation to a zero a of $f(z)$. Then in general a better approximation is $\alpha + \delta\alpha$, where

$$\delta\alpha = -f(\alpha)/f'(\alpha). \qquad (20)$$

This can be seen graphically. In Figure 1 AC is the tangent to the curve $y = f(z)$ at the point $z = \alpha$, and its intersection C with the real axis is the point $z = \alpha + \delta\alpha$.

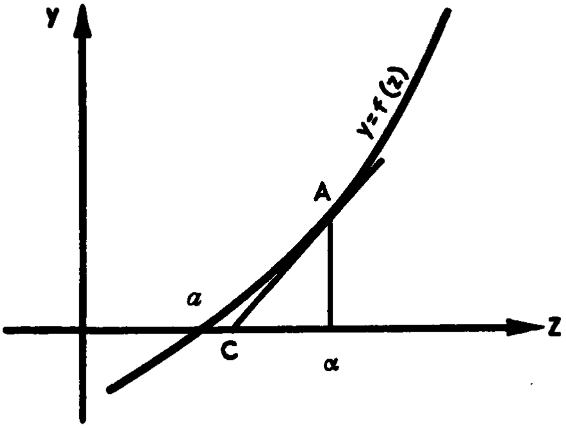


Figure 1

The result (20) can be found by expanding $f(z)$ in the Taylor series centred at $z = \alpha$. We have

$$f(z) = f(\alpha) + (z-\alpha)f'(\alpha) + \frac{1}{2!}(z-\alpha)^2 f''(\alpha) + \ldots = 0, \tag{21}$$

when $z = a$, giving

$$a - \alpha = -\frac{f(\alpha)}{f'(\alpha)} - \frac{1}{2!}\frac{f''(\alpha)}{f'(\alpha)}(a-\alpha)^2 - \ldots. \tag{22}$$

This equation also shows that the error in $\alpha + \delta\alpha$ is of order $(\delta\alpha)^2$. For this reason formula (20) is said to be *quadratically convergent*. The precise implication of (22) is seen on writing it in the form

$$\frac{a-\alpha}{\alpha} = -\frac{f(\alpha)}{\alpha f'(\alpha)} - \frac{1}{2!}\frac{\alpha f''(\alpha)}{f'(\alpha)}\left(\frac{a-\alpha}{\alpha}\right)^2 - \ldots. \tag{23}$$

Clearly if $|\frac{1}{2}\alpha f''(\alpha)/f'(\alpha)|$ is of order unity, the number of correct significant figures in the zero is doubled by each iteration. More generally, if $|\frac{1}{2}\alpha f''(\alpha)/f'(\alpha)|$ is of order 10^k and the number of correct figures in α is S, the number of correct figures in $\alpha + \delta\alpha$ will be $2S - k$.

10. It is of interest to compare Newton's rule with the process of successive linear interpolation described in § 3. Let the errors in z_1, z_2 be respectively h_1, h_2, so that

$$z_1 = a - h_1, \qquad z_2 = a - h_2. \tag{24}$$

Substituting these expressions in the first of (8), and expanding by Taylor's theorem, we find

$$\begin{aligned} z_3 &= a + \frac{h_2 f(a-h_1) - h_1 f(a-h_2)}{f(a-h_2) - f(a-h_1)} \\ &= a + \frac{h_2\{-h_1 f'(a) + \frac{1}{2}h_1^2 f''(a) - \ldots\} - h_1\{-h_2 f'(a) + \frac{1}{2}h_2^2 f''(a) - \ldots\}}{(h_1 - h_2)f'(a) + \ldots} \\ &= a + \frac{f''(a)}{2f'(a)}h_1 h_2 + \ldots. \end{aligned} \tag{25}$$

Therefore if $|\frac{1}{2}af''(a)/f'(a)|$ is of order unity, the relative error in z_3 is the product of the relative errors in z_1 and z_2. In comparison, as we have seen in § 9, each application of Newton's rule squares the relative error in these circumstances.

The total computing effort necessitated by each process is about the same, however. This is because at each step of the successive linear interpolations, only a new function value has to be evaluated, compared with a new function value and a new derivative at each application of Newton's rule.

11. The *order of convergence* of the process of successive linear interpolation may be deduced from the result of the preceding section. Let h_s denote the error in z_s. Then on taking logarithms of (25) we find that for large s

$$\ln h_s = \ln h_{s-1} + \ln h_{s-2}, \tag{26}$$

which is satisfied by $\ln h_s = \rho^s$, if

$$\rho^2 - \rho - 1 = 0.$$

The positive root of this equation, which is the order of convergence, is $\frac{1}{2}(1+\sqrt{5}) = 1{\cdot}62\ldots$.

BAIRSTOW'S PROCESS

12. Newton's rule remains valid if the wanted root is complex. In applying the rule, the easiest way of computing $f(z)$ and $f'(z)$ for a complex value $z = \alpha$ would be to evaluate the remainders on dividing $f(z)$ and $f'(z)$ by the real quadratic polynomial $(z-\alpha)(z-\bar{\alpha})$, where $\bar{\alpha}$ is the conjugate of α. A more convenient iterative process, however, which determines the corrections to the coefficients of an approximate quadratic factor rather than to an approximate root, is due to Bairstow. We first describe the computations involved in dividing $f(z)$ by a quadratic polynomial.

13. Let $z^2 - pz - l$ be a given quadratic polynomial. The division of $f(z)$ by this quadratic is expressed by the equation

$$f(z) = (z^2 - pz - l)\,q(z) + r_1 z + r_0, \tag{27}$$

where the quotient $q(z)$ is a polynomial of degree $n-2$, and $r_1 z + r_0$ is the remainder.

By equating coefficients in (27), we may show that the process of division is equivalent to the application of the recurrence relation

$$q_s = a_s + pq_{s+1} + lq_{s+2}, \tag{28}$$

for $s = n, n-1, \ldots, 0$, starting with $q_{n+2} = q_{n+1} = 0$. Then

$$q(z) = q_n z^{n-2} + q_{n-1} z^{n-3} + \ldots + q_2, \tag{29}$$

$$r_1 = q_1, \qquad r_0 = q_0 - pq_1. \tag{30}$$

14. In Bairstow's process, we first compute the sequence $q_n, q_{n-1}, \ldots, q_0$, and then divide $q(z)$ by $z^2 - pz - l$ by computing the sequence T_{n-2}, $T_{n-1}, \ldots, T_0$, defined by

$$T_s = q_{s+2} + pT_{s+1} + lT_{s+2} \quad (s = n-2, n-3, \ldots, 0), \tag{31}$$

with $T_n = T_{n-1} = 0$. If $z^2 - pz - l$ is an approximate quadratic factor of $f(z)$, a better approximation is given by $z^2 - (p+\delta p)z - (l+\delta l)$, where

$$D\delta l = Mq_1 - T_0 q_0, \qquad D\delta p = T_1 q_0 - T_0 q_1, \tag{32}$$

and

$$M = lT_1 + pT_0, \qquad D = T_0^2 - MT_1. \tag{33}$$

Simple checks on the evaluation of (32) and (33) (but not on the computation of the sequences q_s and T_s) are furnished by the identities

$$T_1 \delta l + T_0 \delta p = -q_1, \qquad T_0 \delta l + M\delta p = -q_0. \tag{34}$$

15. This result may be proved as follows.* If $z^2 - (p+\delta p)z - (l+\delta l)$ were an exact factor, we would have

$$f(z) = \{z^2 - (p+\delta p)z - (l+\delta l)\}\,Q(z), \tag{35}$$

where $Q(z)$ is a polynomial of degree $n-2$.

* Another proof is given in [1].

Subtracting (35) from (27), and using (30), we find

$$(z^2-pz-l)\{q(z)-Q(z)\}+(z\delta p+\delta l)Q(z)+q_1(z-p)+q_0 = 0. \qquad (36)$$

Let α_j ($j = 1, 2$) be the zeros of z^2-pz-l. Then to the first order of small quantities δp and δl, we have $Q(\alpha_j) = q(\alpha_j)$. The value of $q(\alpha_j)$ is obtained from the remainder $T_1 z+T_0-pT_1$ on dividing $q(z)$ by z^2-pz-l; thus

$$q(\alpha_j) = T_1\alpha_j+T_0-pT_1 = T_1(\alpha_j-p)+T_0. \qquad (37)$$

Setting $z = \alpha_j$ in (36) we find that the first term vanishes, and substituting the right of (37) for $Q(\alpha_j)$ we obtain

$$(\alpha_j\,\delta p+\delta l)\{T_1(\alpha_j-p)+T_0\}+q_1(\alpha_j-p)+q_0 = 0.$$

Replacing $\alpha_j(\alpha_j-p)$ by l, we find on reduction

$$(\alpha_j-p)\{T_0\,\delta p+T_1\,\delta l+q_1\}+\{(lT_1+pT_0)\,\delta p+T_0\,\delta l+q_0\} = 0. \qquad (38)$$

Since this equation holds for *both* zeros α_j, we may equate the contents of each set of braces separately to zero. This yields immediately the two equations (34), and solving for δp and δl, we obtain (32).

16. The proof shows that, like Newton's rule, the process is quadratically convergent; the number of correct figures may be doubled by each iteration.

The two formulae are not exactly equivalent, however; this is shown by the fact that Bairstow's process yields an exact result if $f(z)$ is itself a quadratic polynomial, whereas Newton's rule does not.

Example

17. The polynomial

$$f(z) = z^5+0{\cdot}041634z^4+1{\cdot}145460z^3+0{\cdot}163637z^2+0{\cdot}291439z+0{\cdot}142857$$

is divided by the approximate factor

$$z^2-pz-l = z^2-0{\cdot}415z+0{\cdot}487.$$

TABLE 1. BAIRSTOW'S PROCESS

l		−0·487		−0·487280
p		0·415		0·415354
s	a_s	q_s	T_{s-2}	q_s
5	1	1	1	1
4	0·041634	0·456634	0·8716	0·456988
3	1·145460	0·847963	*0·7227*(T_1)	0·847992
2	0·163637	0·293161	*0·1686*(T_0)	0·293173
1	0·291439	*0·000143*(q_1)		*0·000000*(q_1)
0	0·142857	*0·000147*(q_0)		*0·000000*(q_0)
M			−0·2820	
D			0·2322	
δl			−0·000280	
δp			0·000354	

The sequence $q_5, q_4, \ldots, q_0$ given in the centre column of Table 1 is evaluated by application of (28). The next column gives T_3, T_2, T_1, T_0, computed by means of (31). The corrections

$$\delta p = 0{\cdot}000354, \qquad \delta l = -0{\cdot}000280,$$

are found from (32) and (33), and checked by (34). The new factor

$$z^2 - 0{\cdot}415354z + 0{\cdot}487280$$

is checked by division into $f(z)$. Further corrections are obviously less than a unit in the sixth decimal place.

POLYNOMIALS OF HIGH DEGREE

18. The complete solution of polynomial equations

$$f(z) \equiv a_n z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \ldots + a_0 = 0 \tag{39}$$

of degree n exceeding five is seldom worthwhile on desk machines; such problems should be referred, where possible, to organizations equipped with automatic computers. In the remainder of this chapter we describe some of the features of the automatic calculation of the zeros of polynomials with real coefficients. For further details the reader should consult [**36**].

19. An important aspect, which should be fully understood, is that the zeros of a polynomial of high degree, expressed in the explicit form (39), are often poorly determined inasmuch as some, or even all, of them may be extremely sensitive to slight changes in the coefficients.

To determine the effect on a zero α of a small variation in the coefficient a_s, we differentiate (39) with respect to a_s, and then replace z by α. Then

$$f'(\alpha)\frac{\partial\alpha}{\partial a_s} + \alpha^s = 0, \tag{40}$$

giving

$$\frac{\partial\alpha}{\partial a_s} = -\frac{\alpha^s}{f'(\alpha)}, \tag{41}$$

or, in terms of relative errors,

$$\frac{\delta\alpha}{\alpha} = -\frac{\alpha^{s-1}a_s}{f'(\alpha)}\frac{\delta a_s}{a_s}. \tag{42}$$

If $f'(\alpha)$ vanishes, corresponding to a multiple zero, $\partial\alpha/\partial a_s$ is infinite. Thus values of multiple zeros are seriously affected by slight changes in the coefficients. But polynomials with quite distinct zeros may also be afflicted by this malady. Consider, for example,

$$f(z) = (z+1)(z+2)(z+3)\ldots(z+20). \tag{43}$$

One of the worst cases is obtained by taking $\alpha = -16$ and $s = 19$ in (42). We then find that

$$\frac{\delta\alpha}{\alpha} = \frac{16^{18}\times 210}{4!\ 15!}\frac{\delta a_s}{a_s} \doteqdot 3{\cdot}2\times 10^{10}\frac{\delta a_s}{a_s}. \tag{44}$$

This means that in solving a polynomial equation whose roots have a distribution similar to the zeros of (43), we must carry out some parts of the work using at least 10 figures more than are required in the answer, *whatever method of solution is employed.*

General purpose programmes for the solution of polynomial equations must therefore be of high working precision; at least double word-length must be used for most computers.

As a corollary, analogue computers such as isographs and electrolytic tanks, see [**200**], are of little use except for low-degree polynomials, because of their very restricted working precision.

20. The ill-determination of the zeros of high-degree polynomials no doubt suggests to the reader that unless the solution of the physical problem giving rise to a high-degree polynomial equation is itself poorly determined, then another and more stable formulation of the computing problem must exist. This is often true and such a formulation should always be sought. Nevertheless, the polynomial equation, computed and solved with high working precision, may often provide the easiest solution because of the simplicity and generality of its form.

APPLICATION OF ITERATIVE METHODS

21. Among the more suitable methods for automatic work are those based on the use of quadratically convergent iterative formulae with arbitrary initial guesses. Except in pathological cases, repeated applications of Bairstow's process (§ 14) with any initial trial factor will converge after a sufficient number of iterations.

The quadratic factor of $f(z)$ so obtained may be divided out (§ 13), and the process repeated with the quotient replacing $f(z)$. In this way all the factors are obtained. To guard against the possible accumulation of rounding errors, the factors obtained from the successive quotients should subsequently be checked by iteration in the original polynomial equation (39).

22. The process fails when, as might happen in practice, the accumulation of rounding errors is so severe that very poor approximations obtained to the later factors converge on earlier ones when iterated in the original polynomial equation. It can be shown, however, that this will not occur if the zeros of the polynomial are found in ascending order of modulus magnitude; see [**36**].

There is no easy way of ensuring that the zeros are evaluated in precisely this order, but difficulties seldom arise in practice if all the initial approximations to the zeros are taken to be in the neighbourhood of the origin.

23. The number of iterations needed depends on a variety of circumstances, but for a polynomial of degree 20 it is, as a rule, of the order of 10 rather than 100. There is, however, one special situation which should be anticipated. Consider the polynomial $z^{20}-1$, and for illustration take $z = \frac{1}{2}$ as the initial approximation and apply Newton's rule. The next approximation is found to be

$$z = \tfrac{1}{2} + \frac{1-(\frac{1}{2})^{20}}{20(\frac{1}{2})^{19}} \doteq \tfrac{1}{2} + \frac{2^{19}}{20} = 26214{\cdot}9.$$

Further approximations z_r are given by

$$z_{r+1} = z_r + \frac{1-z_r^{20}}{20\,z_r^{19}} \doteq \frac{19}{20}z_r,$$

when z_r is large. The successive approximations therefore diminish slowly and many iterations are required before the neighbourhood of the two real roots ± 1 is approached.

This kind of difficulty may be overcome by imposing the condition

$$|z_{r+1}| \leqslant C|z_r|, \tag{45}$$

where C is an arbitrarily chosen constant, 3 for example; the exact value is not critical. Values of z_{r+1} which violate this inequality are replaced by Cz_r.

A similar safeguard should be used with Bairstow's process.

24. Another device for reducing the number of iterations is to take as the initial approximation at each stage the quadratic factor which has been obtained and divided out at the previous stage. If there is ill-conditioning due to the next factor being close to the previous one, this initial guess will be good. If, however, the factors are well separated, there is no ill-conditioning; the fact that the initial guess is now poor is of no importance. When the first factor obtained has the smallest zeros this procedure results in the other zeros being found in roughly increasing order of magnitude (compare § 22).

25. Advantages of the method described are, first, that iterative processes require relatively few instructions and are easy to programme. Secondly, iterative processes demand the use of no more than the minimum working precision inherently necessary to compensate for ill-conditioning. The root-squaring process, which enjoyed considerable favour on desk machines (see [**37**]), is difficult to programme and suffers more severely from cancellation when applied to ill-conditioned polynomials.

7

FINITE-DIFFERENCE METHODS

DEFINITIONS AND NOTATIONS

1. Table 1 gives four-decimal values of $\sin x$ at 10° intervals of the argument x. In the next column are written down the differences between successive values of $\sin x$; these are called the *first differences*. The next column contains the *second differences*, which are the differences between successive values of the first differences, and so on. It will be noted that the differences steadily decrease in magnitude until they are finally small and oscillatory; a computer would accordingly say that $\sin x$ is 'well-behaved' at this interval. The convention that signs are given at the ends of a column and where they change is a standard one. Tables listed on automatic computers, however, will usually give all the negative signs but not the positive.

TABLE 1

x	$\sin x$						
0°	+0·0000						
		+1736					
10°	·1736		−52				
		1684		−52			
20°	·3420		104		+4		
		1580		48		+0	
30°	·5000		152		4		+4
		1428		44		+4	
40°	·6428		196		8		−7
		1232		36		−3	
50°	·7660		232		5		+6
		1000		31		+3	
60°	·8660		263		8		−1
		737		23		+2	
70°	·9397		286		+10		
		451		−13			
80°	0·9848		−299				
		+152					
90°	+1·0000						

If a general function $y(x)$ is tabulated at equal intervals h, that is, for arguments $x_n = x_0 + nh$, the function $y(x_n)$ may be denoted by y_n. The general scheme of differences may then be set out in three different ways, according to whether the notation of *forward* (Δ), *backward* (∇) or *central* (δ) differences is used. The defining equations are given by

$$\Delta y_n = \nabla y_{n+1} = \delta y_{n+\frac{1}{2}} = y_{n+1} - y_n, \tag{1}$$

and the corresponding tables of differences are shown in Table 2.

It should be noted that the same numerical values are represented in the three schemes: for example, Δ^3_{-1}, ∇^3_2 and $\delta^3_{\frac{1}{2}}$ all represent the same quantity $(y_2 - 3y_1 + 3y_0 - y_{-1})$.

TABLE 2

y_{-2}					y_{-2}					y_{-2}				
	Δ_{-2}					∇_{-1}					$\delta_{-\frac{3}{2}}$			
y_{-1}		Δ^2_{-2}			y_{-1}		∇^2_0			y_{-1}		δ^2_{-1}		
	Δ_{-1}		Δ^3_{-2}			∇_0		∇^3_1			$\delta_{-\frac{1}{2}}$		$\delta^3_{-\frac{1}{2}}$	
y_0		Δ^2_{-1}		etc.	y_0		∇^2_1		etc.	y_0		δ^2_0		etc.
	Δ_0		Δ^3_{-1}			∇_1		∇^3_2			$\delta_{\frac{1}{2}}$		$\delta^3_{\frac{1}{2}}$	
y_1		Δ^2_0			y_1		∇^2_2			y_1		δ^2_1		
	Δ_1					∇_2					$\delta_{\frac{3}{2}}$			
y_2					y_2					y_2				

DETECTION OF ERRORS BY DIFFERENCING

TABLE 3

0		0		0		0
	0		0		0	
0		0		0		+ 1
	0		0		+ 1	
0		0		+1		− 6
	0		+1		− 5	
0		+1		−4		+15
	+1		−3		+10	
1		−2		+6		−20
	−1		+3		−10	
0		+1		−4		+15
	0		−1		+ 5	
0		0		+1		− 6
	0		0		− 1	
0		0		0		+ 1
	0		0		0	
0		0		0		0

2. Table 3 shows how the effect of a single error spreads out fanwise in a table of differences and is at the same time considerably magnified. This fact can be used to detect errors by differencing.

In Table 4 the differences do not decrease smoothly, the terms of δ^4 being bigger than those of δ^3. Comparison with Table 3 suggests that there is an error of -9 in the last place of the entry for $x = 0{\cdot}5$. This is in fact the case; the correct entry is 0·6065 and the last two figures have been transposed, a very common form of error. An error detected in this way should always be corrected by *recomputation* of the function.

It may be mentioned that some accounting machines are very useful for the formation of differences, which can be printed up to the fifth or more at about 300 lines per hour. Punched card machines may also be used and are appreciably faster, if the values are already on cards.

It can easily be verified that the $(n+1)$th differences, like the $(n+1)$th derivative, of an nth degree polynomial, are zero; that is, the nth differences are constant. This makes it possible to build up polynomials from a known constant difference; again accounting machines and punched card machines can be used.

Table 4

x	e^{-x}		δ^2		δ^4
0·0	1·0000				
		−952			
·1	0·9048		+91		
		861		− 9	
·2	·8187		82		+ 1
		779		− 8	
·3	·7408		74		− 8
		705		−16	
·4	·6703		58		+37
		647		+21	
·5	·6056		79		−54
		568		−33	
·6	·5488		46		+36
		522		+ 3	
·7	·4966		49		− 6
		473		− 3	
·8	·4493		+46		
		−427			
0·9	0·4066				

SYMBOLIC RELATIONS

3. Difference formulae are most easily obtained by symbolic methods, regarding the symbols Δ, ∇ and δ as operators. To establish formulae for interpolation, the *displacement operator* E and *averaging operator* μ will also be required. These are defined by the relations

$$Ey_n = y_{n+1}, \qquad \mu y_n = \tfrac{1}{2}(y_{n+\frac{1}{2}} + y_{n-\frac{1}{2}}). \tag{2}$$

The definitions are seen to lead immediately to the relations

$$\left.\begin{aligned} \Delta &= E-1, & \delta &= E^{\frac{1}{2}} - E^{-\frac{1}{2}}, \\ \nabla &= 1-E^{-1}, & \mu &= \tfrac{1}{2}(E^{\frac{1}{2}} + E^{-\frac{1}{2}}), \end{aligned}\right\} \tag{3}$$

from which others may also be obtained. For example,

$$\Delta = \nabla E = \delta E^{\frac{1}{2}}, \qquad \mu^2 = 1 + \tfrac{1}{4}\delta^2. \tag{4}$$

4. To obtain difference formulae for analytical processes such as differentiation and integration it is necessary to establish relations between these operators and the differential operator D, defined by the relation

$$Dy = \frac{dy}{dx}. \tag{5}$$

The connection is provided by Taylor's theorem expressed in the form

$$\begin{aligned} Ey(x) &= y(x+h) \\ &= y(x)+hy'(x)+\frac{h^2}{2!}y''(x)+\ldots \\ &= \left(1+hD+\frac{h^2D^2}{2!}+\ldots\right)y(x) \\ &= e^{hD}y(x), \end{aligned}$$

so that
$$E = e^{hD}. \tag{6}$$

The interrelations between all these operators are summarized in Table 5.

TABLE 5

	E	Δ	δ	∇	hD
E	E	$1+\Delta$	$1+\frac{1}{2}\delta^2+\delta\sqrt{(1+\frac{1}{4}\delta^2)}$	$(1-\nabla)^{-1}$	e^{hD}
Δ	$E-1$	Δ	$\delta\sqrt{(1+\frac{1}{4}\delta^2)}+\frac{1}{2}\delta^2$	$\nabla(1-\nabla)^{-1}$	$e^{hD}-1$
δ	$E^{\frac{1}{2}}-E^{-\frac{1}{2}}$	$\Delta(1+\Delta)^{-\frac{1}{2}}$	δ	$\nabla(1-\nabla)^{-\frac{1}{2}}$	$2\sinh\frac{1}{2}hD$
∇	$1-E^{-1}$	$\Delta(1+\Delta)^{-1}$	$\delta\sqrt{(1+\frac{1}{4}\delta^2)}-\frac{1}{2}\delta^2$	∇	$1-e^{-hD}$
hD	$\log E$	$\log(1+\Delta)$	$2\sinh^{-1}\frac{1}{2}\delta$	$-\log(1-\nabla)$	hD
μ	$\frac{1}{2}(E^{\frac{1}{2}}+E^{-\frac{1}{2}})$	$(1+\frac{1}{2}\Delta)(1+\Delta)^{-\frac{1}{2}}$	$\sqrt{(1+\frac{1}{4}\delta^2)}$	$(1-\frac{1}{2}\nabla)(1-\nabla)^{-\frac{1}{2}}$	$\cosh\frac{1}{2}hD$

INTERPOLATION FORMULAE

5. These formulae express y_p, that is $y(x_0+ph)$, when p is not necessarily an integer, in terms of y_0, y_1 and appropriate differences; the formulae differ in the precise sets of differences employed.

(i) *Newton's interpolation formulae*

6. This formula, using forward differences, is easily obtained in the form

$$\begin{aligned} y_p &= E^p y_0 = (1+\Delta)^p y_0 \\ &= y_0+p\Delta y_0+\frac{p(p-1)}{2!}\Delta^2 y_0+\ldots. \end{aligned} \tag{7}$$

The interpolate obtained by truncating this series at the nth difference is the same as the value taken at x_p by the *interpolating polynomial* of degree n which reproduces exactly the function values at $x_0, x_1, \ldots, x_n$. The corresponding backward-difference formula is obtained in a similar way:

$$\begin{aligned} y_p &= E^p y_0 = (1-\nabla)^{-p} y_0 \\ &= y_0+p\nabla y_0+\frac{p(p+1)}{2!}\nabla^2 y_0+\ldots. \end{aligned} \tag{8}$$

In this case the interpolating polynomial of degree n reproduces the function values at $x_0, x_{-1}, \ldots, x_{-n}$.

These formulae are not satisfactory for use other than near the end of a difference table, when central differences may not be available. By substitution for the forward differences in terms of central differences, Newton's forward-difference formula can be transformed into more suitable formulae, which however may also be obtained directly as follows.

(ii) *Everett's interpolation formula*

7. This formula expresses the interpolate y_p in terms of even differences of y_0 and y_1, so that we assume an expression of the form

$$y_p = (a_0+a_1\delta^2+a_2\delta^4+\dots)y_0+(b_0+b_1\delta^2+b_2\delta^4+\dots)y_1. \qquad (9)$$

Now

$$y_p = E^p y_0 = (1+\Delta)^p y_0, \qquad y_1 = (1+\Delta)y_0, \qquad \delta^2 = \Delta^2(1+\Delta)^{-1}.$$

Thus

$$(1+\Delta)^p y_0 = \Big\{(a_0+b_0+b_0\Delta)+(a_1+b_1+b_1\Delta)\frac{\Delta^2}{1+\Delta} + (a_2+b_2+b_2\Delta)\frac{\Delta^4}{(1+\Delta)^2}+\dots\Big\}y_0. \qquad (10)$$

Multiplying by $(1+\Delta)^s$, expanding the remaining denominators and equating powers of Δ^{2s+1} and Δ^{2s+2}, we obtain the relations

$$b_s = \binom{p+s}{2s+1}, \quad a_{s+1}+b_{s+1} = \binom{p+s}{2s+2}. \qquad (11)$$

Hence

$$a_s = \binom{p+s-1}{2s}-\binom{p+s}{2s+1} = -\binom{p+s-1}{2s+1} = \binom{q+s}{2s+1}, \qquad (12)$$

where $q = 1-p$. Then

$$\begin{aligned} y_p = \quad & qy_0+\frac{(q+1)q(q-1)}{3!}\delta^2 y_0+\frac{(q+2)(q+1)q(q-1)(q-2)}{5!}\delta^4 y_0+\dots \\ & +py_1+\frac{(p+1)p(p-1)}{3!}\delta^2 y_1+\frac{(p+2)(p+1)p(p-1)(p-2)}{5!}\delta^4 y_1+\dots. \end{aligned} \qquad (13)$$

This formula can also be obtained by expressing y_p in the form $y_p = F(q)y_0+F(p)y_1$, where $F(p) = (E^p-E^{-p})/(E-E^{-1})$, and expanding F in powers of δ^2.

We write (13) in the form

$$\begin{aligned} y_p = (1-p)y_0+E_2\delta^2 y_0+E_4\delta^4 y_0+\dots \\ +py_1+F_2\delta^2 y_1+F_4\delta^4 y_1+\dots, \end{aligned} \qquad (14)$$

where E_2, F_2, E_4, F_4, etc., are the *Everett coefficients*. The chief advantage deriving from the use of Everett's formula is that only even-order differences need be tabulated. Also, the interpolating polynomial of degree $2n+1$ reproduces the function values at $x_{n+1}, x_n, \dots, x_{-n}$; since these are centred about the interval (x_0, x_1) the function is thus usually represented more accurately in this interval than it would be by the interpolating polynomial of degree $2n+1$ associated with either (7) or (8).

(iii) *Bessel's interpolation formula*

8. This formula expresses y_p in terms of *mean differences* of even order $\mu\delta^{2n}y_{\frac{1}{2}}$ and odd-order differences $\delta^{2n+1}y_{\frac{1}{2}}$. It is written as

$$y_p = y_0+p\delta y_{\frac{1}{2}}+B_2(\delta^2 y_0+\delta^2 y_1)+B_3\delta^3 y_{\frac{1}{2}}+B_4(\delta^4 y_0+\delta^4 y_1)+\dots, \tag{15}$$

where the *Bessel coefficients* B_2, B_3, etc., are readily obtained with the use of Everett's formula and relations of the form

$$E_2\delta^2 y_0+F_2\delta^2 y_1 = \tfrac{1}{2}(E_2+F_2)(\delta^2 y_0+\delta^2 y_1)+\tfrac{1}{2}(F_2-E_2)\,\delta^3 y_{\frac{1}{2}}, \tag{16}$$

so that $B_2 = \frac{1}{2}(E_2+F_2)$, $B_3 = \frac{1}{2}(F_2-E_2)$, and similarly for coefficients of higher orders.

Bessel's formula is the simplest to use when differences of order greater than the third are negligible.

Examples of the use of interpolation formulae are given in *Interpolation and allied tables* [**167**] and in various text-books. The notation used here is the same as that adopted in this latest edition of *Interpolation and allied tables*, where, in particular, an account will be found of the use of these formulae for *inverse interpolation*, that is, the determination of the value of x_p corresponding to a given value of y.

FORMULAE FOR DERIVATIVES

(i) *Backward- or forward-difference formulae*

9. Formulae giving the derivative at a pivotal point in terms of the backward or forward differences at that point can be obtained immediately from the relations between D, Δ and ∇. With backward differences we find

$$\begin{aligned} hy_0' &= hDy_0 = -\{\log(1-\nabla)\}\,y_0 \\ &= (\nabla+\tfrac{1}{2}\nabla^2+\tfrac{1}{3}\nabla^3+\dots)\,y_0, \end{aligned} \tag{17}$$

and with forward differences we have

$$\begin{aligned} hy_0' &= hDy_0 = \{\log(1+\Delta)\}\,y_0 \\ &= (\Delta-\tfrac{1}{2}\Delta^2+\tfrac{1}{3}\Delta^3-\dots)\,y_0. \end{aligned} \tag{18}$$

The coefficients in these expressions decrease slowly and it is preferable to use central differences if they are available. If a derivative is required at the penultimate point of a table, one of the following formulae may be used to obtain better accuracy:

$$\left.\begin{aligned} hy_0' &= hDE^{-1}y_1 = -(1-\nabla)\{\log(1-\nabla)\}\,y_1 \\ &= (\nabla-\tfrac{1}{2}\nabla^2-\tfrac{1}{6}\nabla^3-\dots)\,y_1, \\ hy_0' &= hDEy_{-1} = (1+\Delta)\{\log(1+\Delta)\}\,y_{-1} \\ &= (\Delta+\tfrac{1}{2}\Delta^2-\tfrac{1}{6}\Delta^3+\dots)\,y_{-1}. \end{aligned}\right\} \tag{19}$$

(ii) *Central-difference formulae*

10. The relation between the second derivative at a pivotal point and the differences centred on that point is obtained immediately in a similar way. We find

$$\begin{aligned} h^2y_0'' &= h^2D^2y_0 = (2\sinh^{-1}\tfrac{1}{2}\delta)^2\,y_0 \\ &= (\delta^2-\tfrac{1}{12}\delta^4+\tfrac{1}{90}\delta^6-\dots)\,y_0. \end{aligned} \tag{20}$$

In the case of the first derivative, however, the series obtained would be in terms of the odd-order differences $\delta^{2n+1}y_0$ which do not actually appear in the difference table. The series required will involve mean differences of odd order $\mu\delta^{2n+1}y_0$ and this is obtained by introducing the factor μ into the numerator, and the corresponding factor $(1+\frac{1}{4}\delta^2)^{\frac{1}{2}}$ into the denominator, in the course of the development. This device is frequently of use. We have

$$\begin{aligned} hy'_0 &= hDy_0 = (2\sinh^{-1}\tfrac{1}{2}\delta)\,y_0 \\ &= (1+\tfrac{1}{4}\delta^2)^{-\frac{1}{2}}(2\sinh^{-1}\tfrac{1}{2}\delta)\,\mu y_0 \\ &= (\mu\delta-\tfrac{1}{6}\mu\delta^3+\tfrac{1}{30}\mu\delta^5-\dots)\,y_0. \end{aligned} \tag{21}$$

Another formula, used in a numerical method for solving first-order differential equations (Chapter 9, § 16), connects the mean of the derivatives y'_0, y'_1 with the mean differences at the half-way point, and is given by

$$\begin{aligned} \tfrac{1}{2}h(y'_0+y'_1) &= h\mu y'_{\frac{1}{2}} = hD\mu y_{\frac{1}{2}} \\ &= (2\sinh^{-1}\tfrac{1}{2}\delta)(1+\tfrac{1}{4}\delta^2)^{\frac{1}{2}}\,y_{\frac{1}{2}} \\ &= (\delta+\tfrac{1}{12}\delta^3-\tfrac{1}{120}\delta^5+\dots)\,y_{\frac{1}{2}}. \end{aligned} \tag{22}$$

Other formulae to suit special circumstances can be obtained by a combination of these methods [167].

FORMULAE FOR NUMERICAL INTEGRATION

(i) *Central-difference formulae*

11. By reversion of the central-difference formula (22) for the mean first derivative at the half-way point, a series is obtained expressing the first difference in terms of mean even differences of the derivative, given by

$$\delta y_{\frac{1}{2}} = (\mu-\tfrac{1}{12}\mu\delta^2+\tfrac{11}{720}\mu\delta^4-\dots)\,hy'_{\frac{1}{2}}. \tag{23}$$

This gives immediately an expression for the integral taken over a single interval, in the form

$$\frac{1}{h}\int_{x_0}^{x_1} y\,dx = (\mu-\tfrac{1}{12}\mu\delta^2+\tfrac{11}{720}\mu\delta^4-\dots)\,y_{\frac{1}{2}}. \tag{24}$$

An integral over an extended range may be evaluated by direct summation of values calculated from this formula; guarding figures may be retained to offset the accumulation of rounding errors. Alternatively, the summation may be performed analytically, giving the formula

$$\begin{aligned} \frac{1}{h}\int_{x_0}^{x_n} y\,dx &= (\tfrac{1}{2}y_0+y_1+\dots+y_{n-1}+\tfrac{1}{2}y_n) \\ &\quad -\tfrac{1}{12}(\mu\delta y_n-\mu\delta y_0)+\tfrac{11}{720}(\mu\delta^3 y_n-\mu\delta^3 y_0)-\dots. \end{aligned} \tag{25}$$

This has the form of the well-known *trapezoidal rule*, together with a difference correction associated with the ends of the range of integration.

12. If a sequence of such integrals is required for successive values of n in the upper limit in the integral of (25), it is convenient to use the first sum δ^{-1}, defined by the relation

$$\delta^{-1}y_{n+\frac{1}{2}}-\delta^{-1}y_{n-\frac{1}{2}} = y_n. \tag{26}$$

The formula for the indefinite integral is obtained immediately as

$$\frac{1}{h}\int^{x_n} y\,dx = (\mu\delta^{-1} - \tfrac{1}{12}\mu\delta + \tfrac{11}{720}\mu\delta^3 - \dots)\,y_n. \qquad (27)$$

To obtain the definite integral from x_0 to x_n, the arbitrary constant in the first sum must be chosen so that the integral vanishes at the lower limit. Thus

$$\delta^{-1}y_{\frac{1}{2}} = \mu\delta^{-1}y_0 + \tfrac{1}{2}y_0 = (\tfrac{1}{2} + \tfrac{1}{12}\mu\delta - \tfrac{11}{720}\mu\delta^3 + \dots)\,y_0. \qquad (28)$$

13. A formula which is used extensively in the integration of second-order differential equations is obtained by reversion of the series (20) for the second derivative, and is given by

$$\delta^2 y_0 = (1 + \tfrac{1}{12}\delta^2 - \tfrac{1}{240}\delta^4 + \dots)\,h^2 y_0''. \qquad (29)$$

(ii) *Backward-difference formulae*

14. Reversion of the two formulae (17) and (18) for the derivative in terms of backward differences leads immediately to the relations

$$\left.\begin{aligned} \nabla y_1 &= (1 + \tfrac{1}{2}\nabla + \tfrac{5}{12}\nabla^2 + \dots)\,hy_0', \\ \nabla y_0 &= (1 - \tfrac{1}{2}\nabla - \tfrac{1}{12}\nabla^2 - \dots)\,hy_0'. \end{aligned}\right\} \qquad (30)$$

The first of these is used as a 'predictor' formula and the second as a 'corrector' in the Adams-Bashforth process for integrating differential equations (Chapter 9, § 8).

(iii) *Gregory's formula*

15. The central-difference formulae should be used wherever possible, since they are the most rapidly convergent. In some cases, however, the integrand may not be readily computable outside the range of integration. In such cases Gregory's formula, which uses only available differences, should be applied. It is obtained by combining the second of (30) with the corresponding expression for Δy_0, given by

$$\Delta y_0 = (1 + \tfrac{1}{2}\Delta - \tfrac{1}{12}\Delta^2 + \dots)\,hy_0'. \qquad (31)$$

We then find Gregory's formula, expressible in the form

$$\begin{aligned} \frac{1}{h}\int_{x_0}^{x_n} y\,dx = {} & (\tfrac{1}{2}y_0 + y_1 + \dots + y_{n-1} + \tfrac{1}{2}y_n) - \tfrac{1}{12}(\nabla y_n - \Delta y_0) \\ & - \tfrac{1}{24}(\nabla^2 y_n + \Delta^2 y_0) - \tfrac{19}{720}(\nabla^3 y_n - \Delta^3 y_0) - \dots. \end{aligned} \qquad (32)$$

(iv) *Simpson's rule*

16. A useful central-difference integration formula is obtained if the integral over two intervals is expressed in terms of the central differences at the middle pivotal point. We find

$$\begin{aligned} \frac{1}{2h}\int_{x_{-1}}^{x_1} y\,dx &= \tfrac{1}{2}\{(hD)^{-1}y_1 - (hD)^{-1}y_{-1}\} \\ &= (hD)^{-1}\mu\delta y_0 \\ &= (2\sinh^{-1}\tfrac{1}{2}\delta)^{-1}(1 + \tfrac{1}{4}\delta^2)^{\frac{1}{2}}\,\delta y_0 \\ &= (1 + \tfrac{1}{6}\delta^2 - \tfrac{1}{180}\delta^4 + \dots)\,y_0. \end{aligned} \qquad (33)$$

In particular, if fourth and higher differences are neglected it can be written in the *Lagrangian* form

$$\int_{x_{-1}}^{x_1} y\,dx \doteq \frac{h}{3}(y_1+4y_0+y_{-1}), \tag{34}$$

which is well known as Simpson's rule.

Bickley [**49**] has listed a number of formulae of this type in which differences above a certain order have been neglected and the rest expressed in terms of pivotal values. These formulae, involving equally-spaced abscissae, are called the *Newton-Cotes* formulae. Unless the differences of y are formed it may be difficult to obtain an accurate estimate of the resulting error. If the accuracy required can be assured, however, this type of formula is valuable and particularly easy to use with automatic computers.

Other integration formulae are considered in Chapter 14.

DIVERGING DIFFERENCES

17. It must be emphasized that the formulae given in this chapter can only be expected to give accurate results for functions which are 'well-behaved' in the sense of § 1. A brief account of the effects of using diverging differences is given by Fox [**76**, page 27].

8

CHEBYSHEV SERIES

INTRODUCTION

1. The formulae of the previous chapter which involve the explicit use of finite differences, are not, in general, well suited to automatic computation. We may recall, for instance, that Table 1 of Chapter 7 gave sufficient information for the computation of $\sin x$ to four decimal places for any x in the range $(0, \frac{1}{2}\pi)$ with the aid of a standard interpolation formula. In conjunction with other finite-difference formulae, it will yield values of integrals and derivatives, though not necessarily to the same accuracy. However, such calculations would require a somewhat elaborate computer programme, and the user of an automatic machine seeks a more convenient procedure.

2. Basically the same information as that given by the above-mentioned table of $\sin x$, is given, to similar accuracy, by the approximate relation

$$\sin \tfrac{1}{2}\pi x \doteqdot 1{\cdot}1336T_1(x) - 0{\cdot}1381T_3(x) + 0{\cdot}0045T_5(x) \quad (-1 \leqslant x \leqslant 1), \qquad (1)$$

where $T_r(x)$ is the *Chebyshev polynomial* of degree r in x, defined by

$$T_r(x) = \cos(r\cos^{-1}x). \qquad (2)$$

The representation of $\sin \frac{1}{2}\pi x$ is thereby achieved with the storage of only three numbers, the coefficients of $T_1(x), T_3(x)$ and $T_5(x)$; the right of (1) may then be readily evaluated, as we shall show later. The expression (1) may also be integrated or differentiated, though again with some qualification regarding accuracy.

The desk-machine user is seldom attracted by the compactness of (1); he usually insists on seeing the function values and differences in order to ascertain at a glance the behaviour of the function, and to obtain a reliable check against isolated computing errors. However, for an automatic computer, which is much less prone to isolated errors, the Chebyshev series representation is preferable.

In this particular example, much of the advantage could also have been gained by use of the approximation

$$\sin \tfrac{1}{2}\pi x \doteqdot 1{\cdot}5708x - 0{\cdot}6460x^3 + 0{\cdot}0797x^5 - 0{\cdot}0047x^7 \quad (-1 \leqslant x \leqslant 1), \qquad (3)$$

obtained by truncating the Taylor-series expansion about $x = 0$. We note that this approximation has one more term than (1); if the Taylor series is truncated after the third term, its maximum error is larger. This is a simple example of the general property that in a given finite range, an approximation in Chebyshev series of prescribed degree represents a function of a real variable more accurately than a truncated Taylor series of the same degree. (In the special case where the function happens to be a polynomial of the required degree, the Chebyshev and Taylor representations are equally accurate, each being a rearrangement of the other.) Moreover, any function which can be represented by an orthodox single-entry table can be represented by a single Chebyshev series, whereas a Taylor series valid over the whole tabular range may not exist.

BEST POLYNOMIAL APPROXIMATION

3. The economy achieved by expansions in Chebyshev series may be regarded as a consequence of the following theorem.

Let $f(x)$ be an arbitrary single-valued function defined in the closed interval (a, b), and suppose $p_n(x)$ to be a polynomial of given degree n such that the deviation $\epsilon_n(x) = f(x) - p_n(x)$ attains its greatest absolute value L at not less than $n+2$ distinct points in (a, b), and is alternately $+L$ and $-L$ at the successive points. Then $p_n(x)$ is the 'best' polynomial approximation of degree n to $f(x)$ in (a, b) in the sense that the maximum value of $|\epsilon_n(x)|$ is as small as possible.

[For let $q_n(x) = f(x) - \eta_n(x)$ be a better polynomial approximation. Then $\epsilon_n(x) - \eta_n(x) = q_n(x) - p_n(x)$ is evidently a polynomial of degree not greater than n which is alternately positive and negative at the $n+2$ (or more) maxima of $|\epsilon_n(x)|$, since at these points $|\eta_n(x)| < |\epsilon_n(x)|$, by hypothesis. This is clearly impossible and so the theorem is proved.]

It should be observed that the theorem does not assert the existence of $p_n(x)$. The conditions are quite reasonable, however, in that they imply a set of $n+2$ equations for L and the $n+1$ coefficients of $p_n(x)$. By similar arguments we can show that if $p_n(x)$ exists, it is unique.

4. The relevance of Chebyshev polynomials to this result is that in the closed interval $(-1, 1)$, the polynomial $T_r(x)$ attains its greatest absolute value (unity) at $r+1$ points, including the end-points, with alternating sign. This is evident from the definition (2), and is illustrated by the diagrams of $T_5(x)$ and $T_6(x)$ opposite which serve as typical examples of odd- and even-order polynomials respectively. The turning values of $T_r(x)$ occur at the points

$$x_s = \cos\frac{s\pi}{r} \quad (s = 0, 1, 2, \ldots, r), \tag{4}$$

and the zeros at

$$x_s = \cos\frac{(s+\frac{1}{2})\pi}{r} \quad (s = 0, 1, 2, \ldots, r-1). \tag{5}$$

This property of $T_r(x)$, together with the theorem of § 3, shows that if $f(x)$ is an arbitrary polynomial of degree $n+1$, the best polynomial approximation of degree n in $(-1, 1)$ is

$$p_n(x) = f(x) - a_{n+1} T_{n+1}(x), \tag{6}$$

where a_{n+1} is a constant chosen so that the coefficient of x^{n+1} on the right of (6) vanishes.

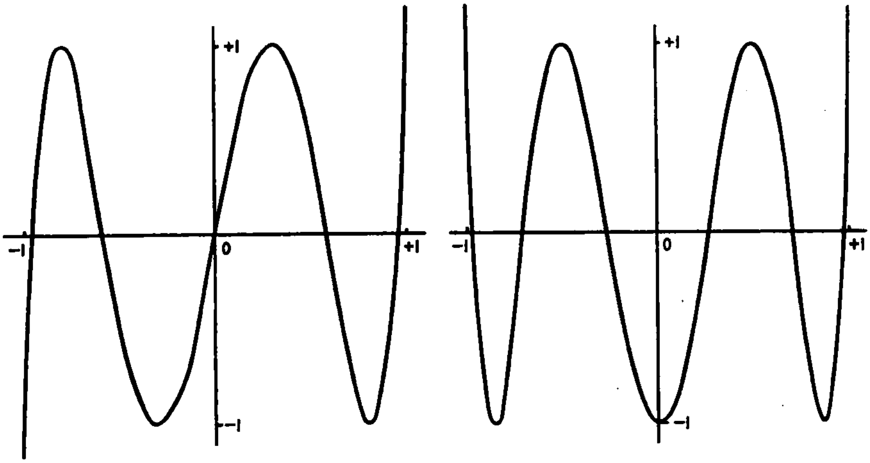
5. No simple explicit expression is known for the best polynomial approximation of given degree to an arbitrary function $f(x)$. Suppose, however, that $f(x)$ can be expanded in the form

$$f(x) = \tfrac{1}{2}a_0 + a_1 T_1(x) + a_2 T_2(x) + \dots \quad (-1 \leqslant x \leqslant 1),$$

which we shall henceforth denote by

$$f(x) = \sum_{r=0}^{\infty}{}' a_r T_r(x) \quad (-1 \leqslant x \leqslant 1), \tag{7}$$

where the prime indicates that the first term of the sum is to be halved.



$T_5(x)$ $T_6(x)$

Figure 1

Then, provided that this series converges reasonably rapidly, the partial sum

$$\sum_{r=0}^{n}{}' a_r T_r(x)$$

will be a good approximation to the best polynomial of degree n in $(-1, 1)$, for the dominant term of the truncation error, whether it be $a_{n+1} T_{n+1}(x)$ or a later term, has the form required by the theorem of § 3. The supposition that the series (7) converges rapidly is often thoroughly justified; indeed Lanczos [**54**] has shown that such expansions are the most strongly convergent of a wide class of expansions in orthogonal polynomials.

CONNEXION WITH FOURIER SERIES

6. The problem of expanding $f(x)$ in a Chebyshev series of the form (7) is essentially the same as that of expanding an arbitrary function in a Fourier cosine series. For if we set $x = \cos\theta$, then (7) becomes

$$f(\cos\theta) = \sum_{r=0}^{\infty}{}' a_r \cos r\theta \quad (0 \leqslant \theta \leqslant \pi). \tag{8}$$

Sufficient conditions for this expansion to exist are that $f(\cos\theta)$ is a continuous function of θ, and has a finite number of maxima and minima in $(0,\pi)$; these are fulfilled if $f(x)$ is a continuous function of x and has a finite number of maxima and minima in $(-1,1)$.

For example, the function $\surd(4x^2+1)$ may be expanded in the form (7), whereas $x\sin(1/x)$ may not. The former illustrates the fact that a function can often be expanded in a Chebyshev series in an interval in which no single Taylor expansion converges.

7. Although an expansion in Chebyshev series is essentially a Fourier cosine series, it has an important property not shared by the general Fourier series. The expansion (7) represents a function $f(\cos\theta)$ which is *naturally* periodic; $f(\cos\theta)$ is at least as well-behaved when regarded as a function of θ in $(-\infty,\infty)$ as is $f(x)$ regarded as a function of x in $(-1,1)$. We may reasonably expect that expansions of such periodic functions in series of trigonometric functions are more rapidly convergent as a class than similar expansions of non-periodic functions, and this is indeed the case.

ADEQUACY OF CHEBYSHEV FORM

8. We may summarize §§ 1 to 7 by restating that a truncated Chebyshev series is normally a good approximation to the best polynomial representation in the sense of § 3. In any given case, the best polynomial approximation of specified degree n may be found by solving the $n+1$ equations obtained by equating to each other, with alternating signs, the $n+2$ expressions for the maximum deviations; the truncated Chebyshev series may be used to provide a first approximation in an iterative procedure. Alternatively, we may use series which give the coefficients in the best polynomial in terms of the coefficients in the infinite Chebyshev series. Hornecker [**58**] has given expressions for the leading terms in such series, and these terms are often sufficient in practice to give the best polynomial to the required accuracy; see also [**59**].

However, it transpires that in practical applications the truncated Chebyshev series is usually very close to the best possible polynomial; the refinements necessary to improve it are seldom worthwhile.

PROPERTIES OF CHEBYSHEV POLYNOMIALS

9. The following properties of Chebyshev polynomials may be derived from the definition (2), as in [**54**].

$$T_{r+1}(x)-2xT_r(x)+T_{r-1}(x)=0, \tag{9}$$

$$\int T_r(x)\,dx=\begin{cases} T_1(x) & (r=0)\\ \tfrac{1}{4}T_2(x) & (r=1)\\ \dfrac{1}{2}\left(\dfrac{T_{r+1}(x)}{r+1}-\dfrac{T_{r-1}(x)}{r-1}\right) & (r>1),\end{cases} \tag{10}$$

$$\int_{-1}^{+1}\frac{T_r(x)\,T_s(x)}{\surd(1-x^2)}\,dx=\begin{cases}\pi & (r=s=0)\\ \tfrac{1}{2}\pi & (r=s\neq 0)\\ 0 & (r\neq s),\end{cases} \tag{11}$$

and, for $n > 0$ and $r, s \leqslant n$,

$$\sum_{j=0}^{n}{}'' T_r(x_j)\, T_s(x_j) = \begin{cases} n & (r = s = 0 \text{ or } n) \\ \tfrac{1}{2}n & (r = s \neq 0 \text{ or } n) \\ 0 & (r \neq s). \end{cases} \tag{12}$$

In (12), $x_j = \cos(\pi j/n)$ and Σ'' denotes a finite sum whose first and last terms are to be halved, so that

$$\sum_{j=0}^{n}{}'' u_j = \tfrac{1}{2}u_0 + u_1 + u_2 + \ldots + u_{n-1} + \tfrac{1}{2}u_n. \tag{13}$$

10. Chebyshev polynomials may be defined for ranges other than $(-1, 1)$, and similar properties derived for them. In practice, however, it is usually more convenient to convert any finite range to $(-1, 1)$ by linear transformation of the variable. An exception to this is the range $(0, 1)$, which is of sufficiently frequent occurrence to merit a notation for its own Chebyshev polynomials, namely

$$T_r^*(x) = T_r(2x-1) = \cos\{r \cos^{-1}(2x-1)\}. \tag{14}$$

Equations similar to those numbered (9) to (12) can be derived without difficulty. We also have

$$T_r^*(x^2) = T_{2r}(x). \tag{15}$$

11. Explicit expressions for the first few Chebyshev polynomials are

$$\begin{aligned} &T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2-1, \quad T_3(x) = 4x^3-3x, \\ &T_4(x) = 8x^4-8x^2+1, \quad T_5(x) = 16x^5-20x^3+5x, \\ &T_6(x) = 32x^6-48x^4+18x^2-1, \end{aligned} \tag{16}$$

and

$$\begin{aligned} &T_0^*(x) = 1, \quad T_1^*(x) = 2x-1, \quad T_2^*(x) = 8x^2-8x+1, \\ &T_3^*(x) = 32x^3-48x^2+18x-1, \quad T_4^*(x) = 128x^4-256x^3+160x^2-32x+1, \\ &T_5^*(x) = 512x^5-1280x^4+1120x^3-400x^2+50x-1. \end{aligned} \tag{17}$$

Similar expressions for higher orders, up to $T_{12}(x)$ and $T_{20}^*(x)$, may be found in [**54**].

CALCULATION OF CHEBYSHEV COEFFICIENTS

12. The coefficients in the Chebyshev expansion of an arbitrary function

$$f(x) = \sum_{r=0}^{\infty}{}' a_r T_r(x) \tag{18}$$

can be obtained in various ways, the most obvious being suggested by the orthogonality relation (11). Thus we have

$$a_r = \frac{2}{\pi}\int_{-1}^{+1} \frac{f(x)\, T_r(x)}{\sqrt{(1-x^2)}}\, dx = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta) \cos r\theta \, d\theta. \tag{19}$$

In practice this formula is seldom used to evaluate the a_r, but it does yield the following upper bounds, discussed in § 17:

$$|a_0| \leqslant 2M, \qquad |a_r| \leqslant \frac{2M}{\pi}\int_0^\pi |\cos r\theta|\, d\theta = \frac{4M}{\pi} \quad (r > 0), \tag{20}$$

where M is the maximum value of $|f(x)|$ in $(-1, 1)$.

13. A more generally useful method of evaluating the coefficients is based on the relation (12). Let us define quantities α_r for $r = 0, 1, 2, \ldots, n$ by

$$\alpha_r = \frac{2}{n}\sum_{j=0}^{n}{}'' f(x_j)\, T_r(x_j) = \frac{2}{n}\sum_{j=0}^{n}{}'' f\left(\cos\frac{\pi j}{n}\right)\cos\frac{\pi r j}{n}, \tag{21}$$

where $x_j = \cos(\pi j/n)$.

If $f(x)$ is a polynomial of degree n or less we have $\alpha_r = a_r$, by virtue of (12). For general $f(x)$, however, this relation is only approximate; with the aid of an obvious extension of (12) we can show that

$$\alpha_r = \frac{2}{n}\sum_{s=0}^{\infty}{}' a_s \sum_{j=0}^{n}{}'' T_s(x_j)\, T_r(x_j) \tag{22}$$

$$= a_r + a_{2n-r} + a_{2n+r} + a_{4n-r} + a_{4n+r} + a_{6n-r} + \ldots . \tag{23}$$

Provided that n is chosen sufficiently large, the coefficients a_s for $s \geqslant n$ and hence the differences $\alpha_r - a_r$ for all r, will vanish to any prescribed accuracy. We then have, within the limits of this accuracy,

$$f(x) = \sum_{r=0}^{n}{}' a_r T_r(x), \tag{24}$$

where

$$a_r = \frac{2}{n}\sum_{j=0}^{n}{}'' f\left(\cos\frac{\pi j}{n}\right)\cos\frac{\pi r j}{n}. \tag{25}$$

14. There is another method for calculating a_r which may be preferable when $f(x)$ satisfies an ordinary linear differential equation; this is described in Chapter 9, §§ 23 to 25.

SUMMATION OF CHEBYSHEV SERIES

15. We now consider the problem of evaluating the Chebyshev series (24), with given numerical coefficients, for an arbitrary value of x in $(-1, 1)$. A simple example of such a series is given in the relation (1). One way is to replace the polynomials $T_r(x)$ by their expressions (16) in powers of x, and then rearrange the result in the form

$$f(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n.$$

Given the coefficients c_r, we may evaluate $f(x)$ for any x in $(-1, 1)$ by the process of nested multiplication (Chapter 6, § 1). Essentially this consists of evaluating successively the quantities $d_n, d_{n-1}, d_{n-2}, \ldots, d_0$ defined by

$$d_r = x d_{r+1} + c_r, \qquad d_{n+1} = 0. \tag{26}$$

The required result is given by

$$f(x) = d_0.$$

16. An alternative procedure avoids this rearrangement; $f(x)$ is evaluated directly from the numerical values of the Chebyshev coefficients a_r by recurrence. We form successively $b_n, b_{n-1}, b_{n-2}, \ldots, b_0$ from

$$b_r = 2xb_{r+1} - b_{r+2} + a_r, \quad b_{n+1} = b_{n+2} = 0. \tag{27}$$

Then

$$f(x) = \tfrac{1}{2}(b_0 - b_2). \tag{28}$$

In [57] it is shown that although the sequence of b_r may be subject to an appreciable building-up error, the resulting error in $f(x)$ is small.

17. For automatic computation, the first of these methods is usually the faster, because its recurrence relation has one term fewer. On the other hand, it suffers from the disadvantage that the coefficients c_r are functions of n; a more complete notation for them would be $c_{n,r}$. A decision to use a lower order of approximation would require the evaluation of a completely fresh set of coefficients, whereas in the second method we merely truncate the series (24) earlier. As a consequence, the publication of 'machine tables' of universal applicability would require a separate set of polynomial coefficients c_r for each n, compared with just one set of Chebyshev coefficients a_r.

A second reason for preferring the Chebyshev series to the rearranged series is that the a_r are bounded, as we saw in (20). The coefficients c_r, in contrast, may become excessively large, thereby restricting the accuracy obtainable with a given word length.

For example, the Bessel function $J_0(10x)$ may be represented in $(-1, 1)$ to nine decimal places by the first thirteen terms of its infinite Chebyshev expansion. The coefficients a_r in this expansion, obtained by using the method of Chapter 9, §§ 23 to 25, are given in Table 1 together with the coefficients c_r in the rearranged polynomial of degree 12 in x^2.

TABLE 1

r	$(-)^r a_r$	$(-)^r c_r$	r	$(-)^r a_r$	$(-)^r c_r$
0	0·06308 1226	1·00000 0000	7	0·00569 8082	240·10331 7504
1	0·21461 6183	24·99999 9868	8	0·00067 7504	93·48251 6480
2	0·00433 6620	156·24999 2936	9	0·00006 0947	28·41391 9232
3	0·26620 3654	434·02762 5984	10	0·00000 4309	6·68205 0560
4	0·30612 5520	678·16663 1936	11	0·00000 0246	1·11987 9168
5	0·13638 8770	678·15586 5600	12	0·00000 0012	0·10066 3296
6	0·03434 7540	470·89281 6384			

We see that although the two expressions are exactly equivalent in that

$$J_0(10x) \doteq \sum_{r=0}^{12}{}' a_r T_{2r}(x) = \sum_{r=0}^{12} c_r x^{2r},$$

there are three more decimal figures (or eleven more binary figures) in the greatest $|c_r|$ than in the greatest $|a_r|$.

18. For evaluating Chebyshev series of the forms

$$f(x) = \sum_{r=0}^{n}{}' a_r T_r^*(x), \tag{29}$$

$$f(x) = \sum_{r=0}^{n}{}' a_r T_{2r}(x), \tag{30}$$

$$f(x) = \sum_{r=0}^{n} a_r T_{2r+1}(x), \tag{31}$$

the necessary modifications of the recurrence method are the replacement of the symbol x on the right of (27) by $2x-1, 2x^2-1, 2x^2-1$ respectively, and in the case of (31) only, the replacement of (28) by

$$f(x) = x(b_0 - b_1). \tag{32}$$

INTEGRATION

19. Given an expansion of the form (7), we may obtain a corresponding expansion for the indefinite integral $\int f(x)\,dx$ by application of (10). Thus

$$\int f(x)\,dx = \text{const.} + \frac{a_0 T_1}{2} + \frac{a_1 T_2}{4} + \sum_{r=2}^{\infty} \frac{a_r}{2}\left(\frac{T_{r+1}(x)}{r+1} - \frac{T_{r-1}(x)}{r-1}\right) = \sum_{r=0}^{\infty}{}' A_r T_r(x), \tag{33}$$

where

$$A_r = \frac{a_{r-1} - a_{r+1}}{2r} \quad (r > 0), \tag{34}$$

and A_0 is determined by the lower limit of integration. As an example we evaluate the expansion for $\int e^x\,dx$, starting with four-decimal values of the coefficients a_r in the expansion of e^x. The analytical formula for the coefficients, obtained from (19), is $a_r = 2I_r(1)$, where I_r is the modified Bessel function of order r.

TABLE 2

r	a_r	A_r	$2I_r(1)$
0	2·5321	—	2·53213
1	1·1303	1·13030	1·13032
2	0·2715	0·27150	0·27150
3	0·0443	0·04433	0·04434
4	0·0055	0·00548	0·00547
5	0·0005	0·00055	0·00054
6	0·0000	0·00004	0·00004

We observe that there is a gain in accuracy in forming the A_r from the a_r, consequent upon the division by $2r$. For further details, including a comparison with other methods for numerical integration, see [**134**].

DIFFERENTIATION

20. The problem of differentiation is the inverse of that of § 19. Given a set $A_0, A_1, A_2, \ldots$, we require the coefficients $a_0, a_1, a_2, \ldots$. They can be found by using (34) in the form

$$a_{r-1} = a_{r+1} + 2r\,A_r. \tag{35}$$

If A_n is the coefficient of highest order which is not negligible, we take $a_n = a_{n+1} = a_{n+2} = \ldots = 0$, and then find $a_{n-1}, a_{n-2}, \ldots, a_0$ by successive application of (35). The factor $2r$ is now multiplicative, and thus gives rise to the loss of accuracy which invariably accompanies numerical differentiation. It is advisable if possible to retain extra decimal places in the coefficients of high order, to minimize this loss.

9

ORDINARY DIFFERENTIAL EQUATIONS: INITIAL VALUE PROBLEMS

INTRODUCTION

1. Ordinary differential equations which arise in practical problems, even when linear and of apparently simple form, can rarely be solved analytically in terms of functions already tabulated. Even if an analytical solution can be found the labour of calculating values of the solution for many values of the independent variable may be considerable. For example, the simple equation

$$\frac{dy}{dx} - \frac{2y}{1-x^4} = 0 \tag{1}$$

has the solution

$$y = A\left(\frac{1+x}{1-x}\right)^{\frac{1}{2}} \exp(\tan^{-1} x), \tag{2}$$

where A is chosen to fit some extra condition such as $y = 1$ when $x = 0$. The systematic tabulation of y for a range of values of x is not a trivial undertaking. In particular, tables are needed of $\tan^{-1} x$ and e^x, and interpolation is necessary in the latter.

The apparently trivial addition of the term x to the right of (1), giving

$$\frac{dy}{dx} - \frac{2y}{1-x^4} = x, \tag{3}$$

increases considerably the complexity of the analytical solution, which is now

$$y = \left(\frac{1+x}{1-x}\right)^{\frac{1}{2}} \exp(\tan^{-1} x)\left\{\int x\left(\frac{1-x}{1+x}\right)^{\frac{1}{2}} \exp(-\tan^{-1} x)\,dx + A\right\}, \tag{4}$$

where A is chosen to fit an extra condition. The evaluation of this expression involves not only the extensive use of mathematical tables but also numerical quadrature.

Other methods of solution, such as the expression of y as a power series

$$y = x^c(a_0 + a_1 x + a_2 x^2 + \ldots), \tag{5}$$

are practicable in only a small number of cases, and may lead to a large amount of work.

2. To overcome these difficulties various numerical methods have been developed. Some are graphical, while some use crude finite-difference approximations to derivatives, and no great precision can be expected from such methods. Most of the methods described in this and the following chapter, however, use formulae which are accurate within prescribed limits, taking into account all significant terms.

The method selected to solve a given differential equation depends largely on the supplementary conditions which specify the particular solution required. These conditions are commonly given either at one or at both of the end-points of the range of integration. For example, in the case of a linear second-order differential equation, two supplementary conditions are normally required to specify a solution. In this case the numerical values of the required solution and of its derivative may be given at one end-point; alternatively the numerical values of the solution may be given at both end-points. The former is an example of an *initial-value* problem, the latter of a *boundary-value* problem; the remainder of this chapter is devoted to methods for solving initial-value problems, boundary-value problems being considered in Chapter 10.

3. When all the supplementary conditions are given at the same point, it is convenient to solve the equation using a step-by-step method, numerical values being calculated at successive pivotal points, equally spaced at an interval h. Most step-by-step methods use finite-difference formulae, but first we discuss a method based on the use of the Taylor series.

THE TAYLOR-SERIES METHOD

4. This method is, in theory, applicable to an equation of any order. As an example, consider the case of the second-order equation

$$y'' = f(x, y, y'). \tag{6}$$

If the function y and its derivative y' have known values y_0, y'_0 at x_0, then y''_0 can be computed directly from the differential equation (6). Moreover differentiation of (6) with respect to x leads to the equation

$$y''' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} y' + \frac{\partial f}{\partial y'} y'', \tag{7}$$

from which y'''_0 can be computed, and so on for higher derivatives. It is then possible to compute the values y_1, y'_1, of y, y' at the next pivotal point $x_1 = x_0 + h$ using the Taylor series

$$y_1 = y_0 + hy'_0 + \frac{h^2}{2!} y''_0 + \frac{h^3}{3!} y'''_0 + \dots, \tag{8}$$

$$y'_1 = \qquad y'_0 + hy''_0 + \frac{h^2}{2!} y'''_0 + \dots. \tag{9}$$

This process may then be repeated in order to advance to the point x_2. First, however, it is usual to check the previous step and calculation of the derivatives by applying the formulae

$$y_0 = y_1 - hy_1' + \frac{h^2}{2!}y_1'' - \frac{h^3}{3!}y_1''' + \dots, \tag{10}$$

$$y_0' = \qquad y_1' - \;hy_1'' \; + \frac{h^2}{2!}y_1''' - \dots. \tag{11}$$

5. In numerical practice it is usual to express the series in terms of *reduced derivatives* $\tau^n \equiv \frac{h^n}{n!}y^{(n)}$, retaining one or two extra figures in the reduced derivatives, and calculating the series in the forms

$$\left.\begin{aligned} y_0 &= (y_1 + \tau_1^2 + \tau_1^4 + \dots) - (\tau_1^1 + \tau_1^3 + \dots), \\ y_2 &= (y_1 + \tau_1^2 + \tau_1^4 + \dots) + (\tau_1^1 + \tau_1^3 + \dots), \\ \tau_0^1 &= (\tau_1^1 + 3\tau_1^3 + 5\tau_1^5 + \dots) - (2\tau_1^2 + 4\tau_1^4 + \dots), \\ \tau_2^1 &= (\tau_1^1 + 3\tau_1^3 + 5\tau_1^5 + \dots) + (2\tau_1^2 + 4\tau_1^4 + \dots). \end{aligned}\right\} \tag{12}$$

6. This method has several advantages. A large interval can often be used; it can be applied, at least theoretically, to non-linear equations; and no special starting procedure is required, an added advantage when an automatic computer is to be used, though the disadvantage of having to use a special starting procedure should not be exaggerated. On the other hand, the derivatives have to be computed, and this may be difficult unless the equation has a simple form such that a recurrence relation for the derivatives can be readily established. For automatic computation it certainly implies extra programming.

The method has been successfully applied on a wide variety of automatic computers, to the equation

$$p(x)y'' + q(x)y' + r(x)y = 0, \tag{13}$$

where p, q and r are quadratic functions of x. In this case there is a five-term recurrence relation between the derivatives.

PREDICTOR-CORRECTOR METHODS

7. These methods are based on finite-difference formulae, but use them in their Lagrangian form. In the case of the first-order equation

$$y' = f(x, y), \tag{14}$$

the method consists of advancing from the point x_n to the point x_{n+1} by means of a quadrature formula which does not include the unknown derivative y'_{n+1}; when the latter has been determined with the aid of the differential equation, the result is corrected by the application of a more accurate formula.

For example, the *predictor* formula

$$y_{n+1} = y_n + \tfrac{1}{12}h(23y'_n - 16y'_{n-1} + 5y'_{n-2}) \qquad (15)$$

may be used to determine an approximate y_{n+1}. An approximate y'_{n+1} is then computed from the differential equation and a more accurate y_{n+1} calculated from the *corrector* formula

$$y_{n+1} = y_n + \tfrac{1}{12}h(5y'_{n+1} + 8y'_n - y'_{n-1}). \qquad (16)$$

From this result y'_{n+1} can be recomputed from (14) and y_{n+1} from (16), and so on until there is no further change.

Similar methods may be applied to equations of higher order. A variety of predictor and corrector formulae have been given by Milne [**63**].

These methods are simple to apply but they are of limited accuracy unless a small interval is used; also they may be unstable (see § 27). From the point of view of automatic computation they suffer from the added disadvantage that a special starting procedure is needed. This will usually mean calculating the first few values with the aid of the Taylor series.

8. The classical method of Adams and Bashforth may be regarded as a predictor-corrector method which makes use of formulae involving backward differences. For first-order equations the formulae are given by

$$y_{n+1} = y_n + (1 + \tfrac{1}{2}\nabla + \tfrac{5}{12}\nabla^2 + \tfrac{3}{8}\nabla^3 + \dots)\,hy'_n \quad \text{(predictor)}, \qquad (17)$$

$$y_{n+1} = y_n + (1 - \tfrac{1}{2}\nabla - \tfrac{1}{12}\nabla^2 - \tfrac{1}{24}\nabla^3 - \dots)\,hy'_{n+1} \quad \text{(corrector)}. \qquad (18)$$

Higher-order equations are solved by repeated application of these equations, though the predictor formula need only be used in the first integration of the derivative of highest order. It may be noted that (17) and (18), truncated after ∇^2, are identical with (15), (16), respectively.

Though there is no truncation error in this method, so that it may be expected to be applicable at a larger interval than those involving Lagrangian formulae, this advantage is offset by the large accumulation of rounding error caused by the slow decrease of the coefficients. For automatic computation this method has little merit; apart from the usual disadvantage of requiring a special starting procedure, the use of a slowly convergent series of differences would make heavy demands on the store.

CENTRAL-DIFFERENCE METHODS

9. Central-difference formulae have more rapidly decreasing coefficients and their use is preferable for accurate work with desk machines. In the methods based on these formulae, in contrast to the methods already described, the quantities used in prediction or extrapolation are not all available when required and must be estimated and afterwards checked. Procedures for first-order and second-order equations are given in some detail.

The first-order equation $y' = f(x, y)$

10. The recorded quantities are y and $2hy'$ and the table shows the situation when values of y_n and $2hy'_n$ have been estimated but not verified

or corrected. Horizontal dashes denote recorded quantities.

x/h	y		δ^2		δ^4	$2hy'$		δ^2		δ^4
$n-4$	–		–		–	–		–		–
		–		–			–		–	
$n-3$	–		–		–	–		–		–
		–		–			–		–	
$n-2$	–		–		–	–		–		–
		–		–			–		–	
$n-1$	–		–			–		–		
		–					–			
n	–					–		×		×
$n+1$										

Computation then proceeds in the following steps:

(i) Estimate $\delta^2(2hy'_n)$ and $\delta^4(2hy'_n)$, denoted by crosses in the table.

(ii) Calculate y_{n+1} from the extrapolation formula

$$y_{n+1} = y_{n-1} + (1 + \tfrac{1}{6}\delta^2 - \tfrac{1}{180}\delta^4 + \dots)\, 2hy'_n. \tag{19}$$

(iii) Compute $2hy'_{n+1}$ from the differential equation, thus correcting the estimate of $\delta^2(2hy'_n)$.

(iv) Calculate $\delta y_{n-\frac{1}{2}}$ from the quadrature formula

$$\delta y_{n-\frac{1}{2}} = \tfrac{1}{2}(\mu - \tfrac{1}{12}\mu\delta^2 + \tfrac{11}{720}\mu\delta^4 - \dots)\, 2hy'_{n-\frac{1}{2}}, \tag{20}$$

thus obtaining a check or correction to the previously extrapolated y_n.

(v) Correct $2hy'_n$ and repeat the cycle if necessary.

The second-order equation $y'' = f(x, y, y')$

11. The recorded quantities are here y, $2hy'$ and $4h^2y''$, and the table shows the situation when tentative values of y_n, $2hy'_n$ and $4h^2y''_n$ have just been recorded.

x/h	y		δ^2		δ^4	$2hy'$		δ^2		δ^4	$4h^2y''$		δ^2		δ^4
$n-4$	–		–		–	–		–		–	–		–		–
		–		–			–		–			–		–	
$n-3$	–		–		–	–		–		–	–		–		–
		–		–			–		–			–		–	
$n-2$	–		–		–	–		–		–	–		–		–
		–		–			–		–			–		–	
$n-1$	–		–			–		–			–		–		
		–					–					–			
n	–					–					–		×		×
$n+1$															

Computation then proceeds in the following steps:

(i) Estimate $\delta^2(4h^2y''_n)$ and $\delta^4(4h^2y''_n)$, shown by crosses in the table.

(ii) Calculate $\delta^2 y_n$, hence building up to an estimate of y_{n+1}, from the formula

$$\delta^2 y_n = \tfrac{1}{4}(1 + \tfrac{1}{12}\delta^2 - \tfrac{1}{240}\delta^4 + \dots)\, 4h^2y''_n. \tag{21}$$

Complete the differencing, obtaining in particular $\delta^3 y_{n-\frac{1}{2}}$ for use in the next step.

(iii) Check or correct the value of $2hy'_{n-1}$, so far only extrapolated, from the formula

$$2hy'_{n-1} = 2(\mu\delta - \tfrac{1}{6}\mu\delta^3 + \tfrac{1}{30}\mu\delta^5 - \dots)\, y_{n-1}. \qquad (22)$$

(iv) Calculate $2hy'_{n+1}$ from the extrapolation formula

$$2hy'_{n+1} = 2hy'_{n-1} + (1 + \tfrac{1}{6}\delta^2 - \tfrac{1}{180}\delta^4 + \dots)\, 4h^2 y''_n. \qquad (23)$$

(v) Calculate $4h^2 y''_{n+1}$ from the differential equation and complete the differencing, correcting the original estimate for $\delta^2(4h^2 y''_n)$. Repeat the cycle if necessary.

12. Though in theory there are no truncation errors it will be noted that a check on the estimated fourth difference is not available until a later stage, and it is advisable to have an interval small enough for the estimation to be performed correctly within a few units. Also it has been assumed that sixth differences are negligible: their inclusion complicates the process.

These methods, or modifications of them, have been very extensively used in the computation of planetary and cometary orbits. They are, however, in many cases inferior to the methods described in the next section and they are not suitable for automatic computation. Apart from the minor disadvantage of requiring an alternative starting procedure, the retention of large numbers of differences makes heavy demands on the store and the estimation process is difficult to programme.

DEFERRED-CORRECTION METHODS

13. The finite-difference methods so far described all use formulae which are properly classed as integration formulae. Another class of methods uses differentiation formulae. Only the function and its differences, but not its derivatives, are recorded. The methods are particularly suitable for linear equations.

The linear second-order equation $y'' + f(x)\,y' + g(x)\,y = k(x)$

14. The derivatives are replaced by central-difference formulae, of which the first terms are expressed in terms of pivotal values and the remainder collected to form the *difference correction* Cy. Then the following equation, the recurrence relation, is obtained and must be satisfied at the pivotal point r:

$$(1 + \tfrac{1}{2}hf_r)\, y_{r+1} - (2 - h^2 g_r)\, y_r + (1 - \tfrac{1}{2}hf_r)\, y_{r-1} + Cy_r = h^2 k_r. \qquad (24)$$

The difference-correction operator is given by

$$C = (-\tfrac{1}{12}\delta^4 + \tfrac{1}{90}\delta^6 - \dots) + hf_r(-\tfrac{1}{6}\mu\delta^3 + \tfrac{1}{30}\mu\delta^5 - \dots). \qquad (25)$$

If two initial values y_0 and y_1 are known, and the difference correction is everywhere ignored, successive pivotal values are calculable from the recurrence relation to form a first approximation $y^{(1)}$. From the differences of $y^{(1)}$ values of $Cy_r^{(1)}$, approximations to those of Cy_r, are calculated and inserted in the recurrence relation from which better approximations to the y_r are obtained. The process is repeated until further changes in y_r are negligible.

After the first use of the recurrence relation, the formula

$$(1+\tfrac{1}{2}hf_r)\eta_{r+1}-(2-h^2g_r)\eta_r+(1-\tfrac{1}{2}hf_r)\eta_{r-1}+Cy_r^{(1)}=0, \quad \eta_0=\eta_1=0 \tag{26}$$

provides *corrections* η to the first approximation $y^{(1)}$. The advantage of calculating the correction rather than the second approximation is that fewer significant figures need be retained.

There is no estimation or truncation and a large interval can be used; the convergence of the method is rapid, more than two cycles rarely being necessary. Though a special starting procedure is required, it is provided by the calculation of y_1 using the Taylor series. The calculation of the first few values of the difference correction requires a knowledge of y_{-1}, y_{-2}, etc., which should be computed by using the recurrence relation in the reverse direction.

The linear equation $y''+f(x)y=k(x)$

15. When the first derivative is absent, the recurrence relation can be written in a form which involves a smaller difference correction, given by

$$\left.\begin{aligned}(1+\tfrac{1}{12}h^2f_{r+1})y_{r+1}-(2-\tfrac{10}{12}h^2f_r)y_r+(1+\tfrac{1}{12}h^2f_{r-1})y_{r-1}+Cy_r\\ =\tfrac{1}{12}h^2(k_{r+1}+10k_r+k_{r-1}),\\ C=\tfrac{1}{240}\delta^6-\tfrac{13}{15120}\delta^8+\dots.\end{aligned}\right\} \tag{27}$$

The linear first-order equation $y'+f(x)y=k(x)$

16. Similar methods can be applied to first-order equations. For example, if the differential equation is used to substitute for y'_{r+1} and y'_r in the formula

$$y_{r+1}-y_r=\tfrac{1}{2}h(y'_{r+1}+y'_r)+(-\tfrac{1}{12}\delta^3+\tfrac{1}{120}\delta^5-\tfrac{1}{840}\delta^7+\dots)y_{r+\frac{1}{2}}, \tag{28}$$

there results the two-term recurrence relation

$$\left.\begin{aligned}y_{r+1}(1+\tfrac{1}{2}hf_{r+1})-y_r(1-\tfrac{1}{2}hf_r)+Cy_{r+\frac{1}{2}}=\tfrac{1}{2}h(k_r+k_{r+1}),\\ C=\tfrac{1}{12}\delta^3-\tfrac{1}{120}\delta^5+\tfrac{1}{840}\delta^7-\dots.\end{aligned}\right\} \tag{29}$$

Details and other applications of these methods have been given by Fox and Goodwin [65].

Non-linear equations

17. In the case of non-linear equations the recurrence relation is also generally non-linear. For example the differential equation

$$y''=f(x,y) \tag{30}$$

can be expressed in the finite-difference form

$$\left.\begin{aligned}\{y_{r+1}-\tfrac{1}{12}h^2f(x_{r+1},y_{r+1})\}-\{2y_r+\tfrac{10}{12}h^2f(x_r,y_r)\}\\ +\{y_{r-1}-\tfrac{1}{12}h^2f(x_{r-1},y_{r-1})\}+Cy_r=0,\\ C=\tfrac{1}{240}\delta^6-\tfrac{13}{15120}\delta^8+\dots.\end{aligned}\right\} \tag{31}$$

The application of Newton's rule (Chapter 6, § 9) for the calculation of y_{r+1} is quick and accurate; the quantities $\partial f_r/\partial y$ used in it also occur in the determination of the correction η which satisfies the *linear* equation

$$\left(1-\frac{1}{12}h^2\frac{\partial f_{r+1}}{\partial y}\right)\eta_{r+1}-\left(2+\frac{10}{12}h^2\frac{\partial f_r}{\partial y}\right)\eta_r+\left(1-\frac{1}{12}h^2\frac{\partial f_{r-1}}{\partial y}\right)\eta_{r-1}+Cy_r^{(1)}=0. \tag{32}$$

Further details and applications of these methods have been given by Clenshaw and Olver [**66**].

18. The particular equation (30), as distinct from those which are non-linear in y', can also be solved by using the recurrence relation

$$\left.\begin{aligned} y_{r+1}-2y_r+y_{r-1}+Cy_r &= h^2f(x_r,y_r),\\ C &= -\tfrac{1}{12}\delta^4+\tfrac{1}{90}\delta^6-\dots\,. \end{aligned}\right\} \tag{33}$$

Though the difference correction is much larger than in (31), this equation is linear in y_{r+1}, the quantity required. The method is easily extended to simultaneous sets of equations of the form (30).

19. These methods are ideally suited for desk machines and they have also been used very successfully on automatic computers. In particular, there are often occasions when they are preferable to many other methods because of their greater stability (see § 27).

THE METHOD OF RUNGE AND KUTTA

20. This method applies to the single first-order equation $y'=f(x,y)$ or to sets of first-order equations; hence it may be used for equations of higher order, which can always be represented as a set of first-order equations. In advancing one step the function $f(x,y)$ is computed at a number of intermediate points, the choice of which is to some extent arbitrary. Those used below are particularly simple and this choice is frequently made.

If the point X has been reached and $y(X) = Y$, we calculate in succession the quantities

$$\left.\begin{aligned} k_0 &= hf(X,Y),\\ k_1 &= hf(X+\tfrac{1}{2}h,\ Y+\tfrac{1}{2}k_0),\\ k_2 &= hf(X+\tfrac{1}{2}h,\ Y+\tfrac{1}{2}k_1),\\ k_3 &= hf(X+h,\ Y+k_2). \end{aligned}\right\} \tag{34}$$

Then $y(X+h)=y(X)+\frac{1}{6}(k_0+2k_1+2k_2+k_3)$ with an error of the order h^5. All processes with an error of this order are called *fourth-order* processes. For other formulae see [**67**], and for further developments [**68**] and [**69**].

Methods of this type are not recommended for desk computation since the frequent calculation of $f(x,y)$ is laborious. They are, however, well suited to automatic computation: no special starting procedure is required; very light demands are made on the store; no estimation is required and a straightforward computational procedure is repeated several times. The calculations are often checked by repetition using a different interval.

21. The extension to the set of first-order equations

$$y_r'=f_r(x,y_1,y_2,\dots,y_n)$$

is immediate and is given by the equations

$$\left.\begin{aligned} k_{r0} &= hf_r(X, Y_1, Y_2, \ldots, Y_n), \\ k_{r1} &= hf_r(X+\tfrac{1}{2}h, Y_1+\tfrac{1}{2}k_{10}, Y_2+\tfrac{1}{2}k_{20}, \ldots, Y_n+\tfrac{1}{2}k_{n0}), \\ k_{r2} &= hf_r(X+\tfrac{1}{2}h, Y_1+\tfrac{1}{2}k_{11}, Y_2+\tfrac{1}{2}k_{21}, \ldots, Y_n+\tfrac{1}{2}k_{n1}), \\ k_{r3} &= hf_r(X+h, Y_1+k_{12}, Y_2+k_{22}, \ldots, Y_n+k_{n2}), \\ y_r(X+h) &= y_r(X)+\tfrac{1}{6}(k_{r0}+2k_{r1}+2k_{r2}+k_{r3}). \end{aligned}\right\} \qquad (35)$$

A disadvantage of the Runge–Kutta process applied to a set of equations is its possible instability (see § 27).

THE METHOD OF DE VOGELAERE

22. This is an interesting hybrid method [70] for solving the second-order equation $y'' = f(x, y)$, in which the first derivative is absent, or a set of such equations. It employs one intermediate point, and the integration from x_r to x_{r+1} is carried out by cyclic use of the equations

$$\left.\begin{aligned} y_{r+\frac{1}{2}} &= y_r+\tfrac{1}{2}hy'_r+\tfrac{1}{24}h^2(4f_r-f_{r-\frac{1}{2}}), \\ y_{r+1} &= y_r+hy'_r+\tfrac{1}{6}h^2(f_r+2f_{r+\frac{1}{2}}), \\ y'_{r+1} &= y'_r+\tfrac{1}{6}h(f_r+4f_{r+\frac{1}{2}}+f_{r+1}), \end{aligned}\right\} \qquad (36)$$

where f_r denotes $f(x_r, y_r)$. The neglected terms are of order h^4, h^5, h^5 respectively and the method, is, in fact, comparable in accuracy with the fourth-order Runge–Kutta process. The function f is, however, computed only twice per step.

Though at the start it is necessary to know not only y_0, y'_0 but also $f_{-\frac{1}{2}}$, this quantity is readily obtained from $y_{-\frac{1}{2}}$, given to sufficient accuracy by

$$y_{-\frac{1}{2}} = y_0-\tfrac{1}{2}hy'_0+\tfrac{1}{8}h^2f_0. \qquad (37)$$

SOLUTION IN CHEBYSHEV SERIES

23. Two methods which take advantage of the properties of Chebyshev polynomials have been proposed for linear equations whose coefficients are polynomials in x. Lanczos [6] ('the τ-method') finds the coefficients of a polynomial solution of the differential equation perturbed by a small multiple τ of $T_n(x)$, while Clenshaw [71] calculates directly the coefficients of the Chebyshev-series expansion of the solution. The latter method, described below, is often more convenient in practice.

Suppose the range of integration is normalized to $-1 \leqslant x \leqslant 1$. We assume the expansion

$$y(x) = \tfrac{1}{2}a_0+a_1T_1(x)+a_2T_2(x)+\ldots . \qquad (38)$$

Assume similar expansions for the derivatives

$$y^{(s)}(x) = \tfrac{1}{2}a_0^{(s)}+a_1^{(s)}T_1(x)+a_2^{(s)}T_2(x)+\ldots \quad (s = 1, 2, \ldots, q), \qquad (39)$$

where q is the order of the differential equation. Then from the relation

$$2\int T_r(x)\,dx = \frac{T_{r+1}(x)}{r+1}-\frac{T_{r-1}(x)}{r-1} \qquad (40)$$

we obtain

$$2ra_r^{(s)} = a_{r-1}^{(s+1)} - a_{r+1}^{(s+1)} \quad (r \geqslant 1), \tag{41}$$

a relation which holds for $s = 0$ if we define $a_r^{(0)}$ to be a_r.

If, in addition, $C_r(f)$ is used to denote the coefficient of T_r in the expansion of a function f, and *twice* this coefficient when $r = 0$, then from the relation

$$2xT_r(x) = T_{|r-1|}(x) + T_{r+1}(x) \quad (r = 0, 1, 2, \ldots), \tag{42}$$

we obtain

$$C_r(xy^{(s)}) = \tfrac{1}{2}(a_{|r-1|}^{(s)} + a_{r+1}^{(s)}) \quad \begin{pmatrix} r = 0, 1, 2, \ldots \\ s = 1, 2, \ldots, q \end{pmatrix}, \tag{43}$$

and hence

$$C_r(x^p\, y^{(s)}) = \frac{1}{2^p} \sum_{j=0}^{p} \binom{p}{j} a_{|r-p+2j|}^{(s)} \quad \begin{pmatrix} r = 0, 1, 2, \ldots \\ s = 1, 2, \ldots, q \end{pmatrix}. \tag{44}$$

If the series (38), (39) are substituted into the differential equation, then, using the relations (41), (44) we obtain an infinite set of simultaneous equations for the coefficients $a_r^{(s)}$. These are solved by recurrence, with the assumption that $a_r^{(s)} = 0$ for r greater than some suitably large N, to determine the coefficients $a_r^{(s)}$ $(r = N-1, N-2, \ldots)$ corresponding to one or more sets of assumed values of the $a_N^{(s)}$ $(s = 1, 2, \ldots, q)$. The solutions so obtained are then combined to satisfy the initial conditions.

24. As an example, consider the solution of the equation

$$xy'' + y' + 16xy = 0 \tag{45}$$

in the range $0 \leqslant x \leqslant 1$, with initial conditions

$$y(0) = 1, \qquad y'(0) = 0. \tag{46}$$

This corresponds to the solution of Bessel's equation for $J_0(x)$ over the range 0 to 4.

The solution is an even function of x and so the T_r of odd order do not appear in its expansion. Thus, from (45),

$$C_r(xy'') + C_r(y') + 16C_r(xy) = 0 \quad (r = 1, 3, 5, \ldots), \tag{47}$$

and using (43), we find that

$$\tfrac{1}{2}(a''_{r+1} + a''_{r-1}) + a'_r + 8(a_{r+1} + a_{r-1}) = 0 \quad (r = 1, 3, 5, \ldots). \tag{48}$$

Equations (48) and (41) could now be used to compute the coefficients. However, it is simpler first to use (41) to eliminate the a''_r. Equation (48) implies that

$$\begin{aligned} &\tfrac{1}{2}(a''_{r-2} + a''_r - a''_r - a''_{r+2}) + (a'_{r-1} - a'_{r+1}) \\ &\qquad + 8(a_{r-2} + a_r - a_r - a_{r+2}) = 0 \quad (r = 2, 4, \ldots), \end{aligned} \tag{49}$$

whence, using (41), we obtain

$$r(a'_{r-1} + a'_{r+1}) + 8(a_{r-2} - a_{r+2}) = 0 \quad (r = 2, 4, \ldots). \tag{50}$$

Equations (41) and (50) are then used alternately in the recurrence process, in the forms

$$\left.\begin{aligned} a'_{r-1} &= a'_{r+1} + 2ra_r \\ a_{r-2} &= a_{r+2} - \tfrac{1}{8}r(a'_{r-1} + a'_{r+1}) \end{aligned}\right\} \quad (r = N, N-2, \ldots, 2). \tag{51}$$

25. In a typical case we take $a_{10} = 1$ with all higher order coefficients zero; this leads to the trial solution given in Table 1.

TABLE 1

r	Trial Values		Final a_r
	a_r	a'_{r+1}	
0	-1089	$+11220$	$+0{\cdot}1003$
2	$+7225$	-17680	$-0{\cdot}6653$
4	-2704	$+\ 3952$	$+0{\cdot}2490$
6	$+\ 361$	$-\ 380$	$-0{\cdot}0332$
8	$-\ 25$	$+\ 20$	$+0{\cdot}0023$
10	$+\ 1$		$-0{\cdot}0001$

The second condition (46) has been satisfied automatically and only the condition $y(0) = 1$ remains. This is satisfied by dividing the trial solution by the constant factor

$$(\tfrac{1}{2}a_0 - a_2 + a_4 - a_6 + \ldots) = -10860{\cdot}5,$$

leading to the final values of a_r given in the table. As a check, the sum of the series at $x = 1$ is $-0{\cdot}3972$, in agreement with the true value $J_0(4) = -0{\cdot}39715\ldots$.

The precision of the results may always be increased by taking a larger value of N.

BUILDING-UP ERRORS

26. Suppose an attempt is made to obtain the solution $y = e^{-x}$ to the equation $y'' - y = 0$ with the initial conditions $y(0) = 1, y'(0) = -1$. Then, however many significant figures are kept in the computation, the rounding error will introduce a small multiple of the unwanted solution e^x and, if the computation is carried far enough, this solution will increase to such an extent that it will eventually swamp the required solution.

This phenomenon is known as *building-up error*. In the case given it could easily be avoided by computing in the reverse direction from known values of e^{-x} and its derivative for a large value X of x. In general, however, such a simple procedure will not be available, and some ingenuity is needed to obtain an accurate solution. Considerable experience is required and this section is included merely as a warning.

STABILITY

27. In the previous subsection we have mentioned the difficulty of obtaining a decreasing solution of a differential equation in the presence of an unwanted increasing solution. Sometimes, although the original differential equation does not have an unwanted increasing solution, the associated finite-difference equation does have such a solution. In such a case we say that the method is *unstable*.

Instability of this nature may arise in two ways:

(i) The finite-difference equation may be of a higher order than the differential equation it represents, and an additional solution so introduced may be increasing.

(ii) If the differential equation has some solutions which decrease very rapidly compared with the others, then it may happen that only the latter are adequately represented by the finite-difference equation, while the former are transformed into rapidly increasing functions.

28. As an example of the first type of instability, let us attempt to solve the system

$$y' = -\lambda y, \qquad y(0) = 1, \qquad (\lambda > 0), \tag{52}$$

which has the solution $y = e^{-\lambda x}$, by any predictor-corrector method which uses Simpson's rule as a corrector [**63**]. Then

$$y_{n+1} - y_{n-1} = \tfrac{1}{3}h(y'_{n+1} + 4y'_n + y'_{n-1}). \tag{53}$$

From (52) and (53) we obtain

$$(1 + \tfrac{1}{3}\lambda h)\, y_{n+1} + \tfrac{4}{3}\lambda h y_n - (1 - \tfrac{1}{3}\lambda h)\, y_{n-1} = 0. \tag{54}$$

The general solution of this difference equation is

$$y_n = AE_1^n + BE_2^n, \tag{55}$$

where A, B are arbitrary constants and

$$E_1, E_2 = [-\tfrac{2}{3}\lambda h \pm \sqrt{(1 + \tfrac{1}{3}\lambda^2 h^2)}]/(1 + \tfrac{1}{3}\lambda h). \tag{56}$$

The ratio E of consecutive values y_{n+1}, y_n of the true solution is $e^{-\lambda h}$, and on expansion E_1 is found to differ from this only in terms of the fifth and higher orders. On the other hand $|E_2|$ is always greater than unity, so that an unwanted increasing solution has been introduced and the method is in fact unstable.

29. As an illustration of the second type of instability, consider the equations

$$y' = -10y + 6z, \qquad z' = 13{\cdot}5y - 10z, \tag{57}$$

with $y(0) = \tfrac{4}{3}e, z(0) = 0$. The analytic solution is

$$y = \tfrac{2}{3}e(e^{-x} + e^{-19x}), \qquad z = e(e^{-x} - e^{-19x}). \tag{58}$$

For values of x greater than unity the second exponential is negligible to seven decimal places and it might be expected that the equations could be integrated at an interval of, say, $h = 0{\cdot}2$. In fact, if we apply the Runge–Kutta process (35), starting from $x = 1$ and retaining two decimals, we obtain the results of Table 2, which are clearly incorrect.

TABLE 2

x	1·0	1·2	1·4	1·6	1·8	2·0
y	0·67	0·55	0·46	0·40	0·41	0·68
z	1·00	0·82	0·66	0·51	0·29	−0·28

The explanation lies in the fact that when the Runge–Kutta process (34) is used to integrate $y' = -\lambda y$, it represents $E = e^{-\lambda h}$ by

$$E_1 = 1 - \lambda h + \tfrac{1}{2}\lambda^2 h^2 - \tfrac{1}{6}\lambda^3 h^3 + \tfrac{1}{24}\lambda^4 h^4.$$

This is a good approximation for the first exponential, for which $\lambda h = 0{\cdot}2$ but for the second $\lambda h = 3{\cdot}8$, so that $e^{-3\cdot 8} = 0{\cdot}02\ldots$ is replaced by $E_1 = 3{\cdot}96\ldots$. In fact this process is stable if $\lambda h < 2{\cdot}8$, since $|E_1| < 1$ for this range.

30. With a large set of simultaneous equations, which in general will not have constant coefficients and may well be non-linear, it is often difficult to determine whether or not rapidly decreasing functions of this type will be present. It is essential that the method used is stable for the interval selected. This requirement would sometimes necessitate the use of an extremely small interval with the Runge–Kutta process, and other methods are then preferable. For example, formula (29), with the difference correction neglected, applied to the equation $y' = -\lambda y$, leads to the approximation

$$E_1 = \frac{1 - \frac{1}{2}\lambda h}{1 + \frac{1}{2}\lambda h}$$

and $|E_1|$ is less than unity however large λ. This method is accordingly stable and it could, for example, be used to integrate the equations (57) quite satisfactorily at the interval $h = 0{\cdot}2$.

A full discussion of stability problems is outside the scope of this manual. As a general rule it is advisable to test any method on simple equations such as $y' = -\lambda y$. If a method that may be unstable has to be used, particular care must be taken to ensure that the results are subject to adequate independent checks.

10

ORDINARY DIFFERENTIAL EQUATIONS: BOUNDARY-VALUE PROBLEMS

INTRODUCTION

1. In this chapter we consider the solution of differential equations for which the supplementary conditions are given at both ends of the range of integration. For illustration we use the linear second-order equation

$$y'' + f(x)\,y' + g(x)\,y = k(x), \tag{1}$$

though the principal features of the methods apply equally to linear equations of higher order, and to systems of linear differential equations.

DIRECT FINITE-DIFFERENCE METHOD

2. As in the deferred-correction method, described in Chapter 9, § 14, we replace each derivative by a central-difference formula, of which the first term is expressed in terms of pivotal values and the remaining terms constitute the difference correction Cy. This leads to the system of equations

$$(1-\tfrac{1}{2}hf_r)\,y_{r-1} - (2-h^2 g_r)\,y_r + (1+\tfrac{1}{2}hf_r)\,y_{r+1} + Cy_r = h^2 k_r \qquad (r = 1, 2, \ldots, n-1), \tag{2}$$

where

$$C = (-\tfrac{1}{12}\delta^4 + \tfrac{1}{90}\delta^6 - \ldots) + hf_r(-\tfrac{1}{6}\mu\delta^3 + \tfrac{1}{30}\mu\delta^5 - \ldots). \tag{3}$$

If we disregard the difference corrections there are $n-1$ equations in the $n+1$ unknowns $y_0, y_1, y_2, \ldots, y_{n-1}, y_n$.

3. The extra two equations needed to provide a solution are obtained from the boundary conditions. The simplest and probably most common case is that in which both y_0 and y_n are known. In this case we merely substitute their values in the $n-1$ equations, which then can be written in the form

$$\left.\begin{aligned}
(2-h^2 g_1)\,y_1 + (1+\tfrac{1}{2}hf_1)\,y_2 &= h^2 k_1 - Cy_1 - (1-\tfrac{1}{2}hf_1)\,y_0,\\
(1-\tfrac{1}{2}hf_2)\,y_1 - (2-h^2 g_2)\,y_2 + (1+\tfrac{1}{2}hf_2)\,y_3 &= h^2 k_2 - Cy_2,\\
(1-\tfrac{1}{2}hf_3)\,y_2 - (2-h^2 g_3)\,y_3 + (1+\tfrac{1}{2}hf_3)\,y_4 &= h^2 k_3 - Cy_3,\\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots &\\
(1-\tfrac{1}{2}hf_{n-2})\,y_{n-3} - (2-h^2 g_{n-2})\,y_{n-2} + (1+\tfrac{1}{2}hf_{n-2})\,y_{n-1} &= h^2 k_{n-2} - Cy_{n-2},\\
(1-\tfrac{1}{2}hf_{n-1})\,y_{n-2} - (2-h^2 g_{n-1})\,y_{n-1} &= h^2 k_{n-1} - Cy_{n-1}\\
&\qquad - (1+\tfrac{1}{2}hf_{n-1})\,y_n,
\end{aligned}\right\} \tag{4}$$

or in shortened matrix notation as

$$\mathbf{A}\mathbf{y} = \mathbf{d} - \mathbf{C}\mathbf{y}. \tag{5}$$

Here $\mathbf{C}\mathbf{y}$ denotes the vector of difference corrections Cy_r, $\mathbf{d}$ the remaining terms on the right-hand side, which are known, and $\mathbf{A}$ is a square matrix of special form, given by

$$\mathbf{A} = \begin{bmatrix} b_1 & c_1 & & & & & \\ a_2 & b_2 & c_2 & & & & \\ & a_3 & b_3 & c_3 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & & a_{n-2} & b_{n-2} & c_{n-2} \\ & & & & & a_{n-1} & b_{n-1} \end{bmatrix}. \tag{6}$$

All the elements are zero except those in the main diagonal and two adjacent diagonals. This is sometimes called a *band matrix of width three* or a *triple-diagonal matrix* (compare Chapter 3, § 16).

4. In (5) the term $\mathbf{C}\mathbf{y}$ is not included in $\mathbf{d}$ because its values are not known in advance, but depend on the y values, so far unknown. The difference correction depends on the size of the interval h, and will be negligible for sufficiently small h. Even for much larger h its values may not be very large, so that a suitable method is one of successive approximation, and proceeds as follows.

(i) In equation (5) a first approximate solution $y^{(1)}$ is obtained by neglecting $\mathbf{C}\mathbf{y}$ and solving the equations

$$\mathbf{A}\mathbf{y}^{(1)} = \mathbf{d}. \tag{7}$$

The difference between $y^{(1)}$ and the true solution y depends on the size of $\mathbf{C}\mathbf{y}$ and therefore on the interval h.

(ii) We next calculate a correction η_1 by differencing the $y^{(1)}$ obtained from (7), calculating $\mathbf{C}\mathbf{y}^{(1)}$ (which will be a close approximation to $\mathbf{C}\mathbf{y}$) and solving the equations

$$\mathbf{A}\eta_1 = -\mathbf{C}\mathbf{y}^{(1)}. \tag{8}$$

The boundary values of η_1 are of course zero, since $y^{(1)}$ already has its correct values at these points, and the terms $h^2 k_r$ have been included in the calculation of $y^{(1)}$, so that the right of (8) contains only the difference correction, and the matrix on the left of (8) is identical with (6).

(iii) The process can be continued: if $\mathbf{C}\eta_1$ is significant we can calculate a further correction η_2 from an equation corresponding to (8), given by

$$\mathbf{A}\eta_2 = -\mathbf{C}\eta_1, \tag{9}$$

and repeat until there is no further change. In practice the cycle rarely needs to be performed more than twice.

5. The calculation of the difference correction, using central differences as in the expression (3), cannot be performed immediately at points near

the ends of the range, since central differences do not exist there. It is clear from equation (2), however, that if any two adjacent values y_r, y_{r+1} are given, we can calculate y_{r-1} and y_{r+2} directly, writing (2) in the respective forms

$$\left.\begin{aligned}(1-\tfrac{1}{2}hf_r)\,y_{r-1} &= (2-h^2g_r)\,y_r \quad\;\; -(1+\tfrac{1}{2}hf_r)\,y_{r+1}+h^2k_r \;\; -Cy_r,\\ (1+\tfrac{1}{2}hf_{r+1})\,y_{r+2} &= (2-h^2g_{r+1})\,y_{r+1}-(1-\tfrac{1}{2}hf_{r+1})\,y_r+h^2k_{r+1}-Cy_{r+1}.\end{aligned}\right\}\qquad(10)$$

The terms **Cy** are of course neglected in the first approximation and included in others. Equations similar to (10) can be used to extend the solution at each end of the range x_0 to x_n, providing material from which central differences can be obtained for use in **Cy**.

6. The choice of interval h is governed by two factors. First, if h is sufficiently small for the difference correction to be negligible, then the matrix **A** may be of very large order, the labour of calculation excessive, and high accuracy difficult to achieve. Second, the interval must not be so large that the differences do not converge, since the finite-difference equations then have no meaning: very slow convergence of the differences may also be inconvenient, involving many repetitions of the iterative process. These restrictions apart, we can choose any value of h, and it is usually possible to keep to an acceptable size the number of linear equations involved.

7. As an example we consider the solution of the simple equation

$$y''+y = 0, \qquad (11)$$

with $y(0) = 0, y(1) = 0{\cdot}84147$. Equations (2) and (3) become

$$\left.\begin{aligned}&y_{r-1}-(2-h^2)\,y_r+y_{r+1}+Cy_r = 0,\\ &C = -\tfrac{1}{12}\delta^4+\tfrac{1}{90}\delta^6-\ldots,\end{aligned}\right\}\qquad(12)$$

and, with interval $h = 0{\cdot}2$, the basic equations are

$$y_{r-1}-1{\cdot}96y_r+y_{r+1}+Cy_r = 0. \qquad (13)$$

The complete set of equations corresponding to (4) is given by

$$\left.\begin{aligned}-1{\cdot}96y(0{\cdot}2)+\quad y(0{\cdot}4)\qquad\qquad\qquad\qquad\qquad\qquad &= -Cy(0{\cdot}2),\\ y(0{\cdot}2)-1{\cdot}96y(0{\cdot}4)+\quad y(0{\cdot}6)\qquad\qquad\qquad &= -Cy(0{\cdot}4),\\ y(0{\cdot}4)-1{\cdot}96y(0{\cdot}6)+\quad y(0{\cdot}8) &= -Cy(0{\cdot}6),\\ y(0{\cdot}6)-1{\cdot}96y(0{\cdot}8) &= -Cy(0{\cdot}8)-0{\cdot}84147.\end{aligned}\right\}\qquad(14)$$

The first approximation $y^{(1)}$, obtained by neglecting **Cy** in (14), has the values

$y(0{\cdot}2) = 0{\cdot}19878,\quad y(0{\cdot}4) = 0{\cdot}38962,\quad y(0{\cdot}6) = 0{\cdot}56486,\quad y(0{\cdot}8) = 0{\cdot}71752,$

and the use of (13), with Cy_r neglected, to obtain external values to the same degree of approximation, gives the results shown in Table 1.

TABLE 1

x	y		δ^2		δ^4		δ^6	Cy
−0·4	−0·38961							
		+19083						
−0·2	−0·19878		+ 795					
		19878		−795				
0·0	0·00000		0		+ 1			
		19878		794		+27		
+0·2	+0·19878		− 794		28		+13	−2·3
		19084		766		40		
0·4	0·38962		1560		68		−23	−5·7
		17524		698		17		
0·6	0·56486		2258		85		+16	−7·1
		15266		613		33		
0·8	0·71752		2871		118		−17	−9·8
		12395		495		+16		
1·0	0·84147		3366		+134			
		9029		−361				
1·2	0·93176		−3727					
		+ 5302						
+1·4	+0·98478							

The sixth differences are oscillating about zero so that we ignore them in the calculation of the difference correction; this is then just $-\frac{1}{12}\delta^4 y$ and is given in the table, with an extra figure, opposite the relevant pivotal points. Insertion of these results in the correction equations, given by (14) with the exclusion of the constant −0·84147, gives the corrections

$$\eta(0{\cdot}2) = -0{\cdot}00011, \quad \eta(0{\cdot}4) = -0{\cdot}00020,$$
$$\eta(0{\cdot}6) = -0{\cdot}00022, \quad \eta(0{\cdot}8) = -0{\cdot}00016.$$

When we difference these quantities, as shown in Table 2, there is clearly no further correction, and the final solution is $y^{(1)} + \eta$, given by

$$y(0{\cdot}2) = 0{\cdot}19867, \quad y(0{\cdot}4) = 0{\cdot}38942,$$
$$y(0{\cdot}6) = 0{\cdot}56464, \quad y(0{\cdot}8) = 0{\cdot}71736,$$

agreeing with the analytical solution $y = \sin x$.

TABLE 2

x	η		δ^2		δ^4
+0·0	−0·00000				
		−11			
0·2	0·00011		+ 2		
		− 9		+5	
0·4	0·00020		+ 7		−4
		− 2		+1	
0·6	0·00022		+ 8		+1
		+ 6		+2	
0·8	0·00016		+10		
		+16			
+1·0	−0·00000				

8. If some other boundary condition is imposed, involving the first derivative, a slightly different procedure is necessary at that boundary.

If the boundary is at x_0, we satisfy the differential equation at this point also by using the equation

$$(1-\tfrac{1}{2}hf_0)y_{-1}-(2-h^2g_0)y_0+(1+\tfrac{1}{2}hf_0)y_1+Cy_0 = h^2k_0, \qquad (15)$$

which is equation (2) with $r = 0$. The boundary condition will have the form

$$y_0'+ay_0 = b, \qquad (16)$$

which we replace by its central-difference equivalent

$$y_1-y_{-1}+2hay_0+C_1y_0 = 2hb, \qquad (17)$$

where C_1 is a new difference-correction operator, given by

$$C_1 = 2(-\tfrac{1}{6}\mu\delta^3+\tfrac{1}{30}\mu\delta^5-\ldots), \qquad (18)$$

and the first difference $\mu\delta y_0$ in the derivative formula has been replaced by $\frac{1}{2}(y_1-y_{-1})$. The point x_{-1} is external to the range, and we eliminate the value y_{-1} from (15) with the use of (17), giving the new equation

$$\begin{aligned}\{(1-\tfrac{1}{2}hf_0)\,2ha-(2-h^2g_0)\}\,y_0+2y_1 \\ = h^2k_0+2hb(1-\tfrac{1}{2}hf_0)-Cy_0-(1-\tfrac{1}{2}hf_0)\,C_1y_0. \qquad (19)\end{aligned}$$

This is the first of the new equations corresponding to (4), and the second is obtained by moving the term in y_0 over to the left in the first of (4). The remaining equations are unchanged until the second boundary is reached, where the procedure again depends on the boundary condition. In the worst case, in which a first derivative occurs at both boundaries, the new matrix will have order $n+1$, corresponding to the unknowns $y_0, y_1, \ldots, y_{n-1}, y_n$, but it will still be a band matrix like (6).

For the calculation of the difference corrections, including that of type C_1y_0 in (19), values external to the range can be calculated as before, once approximations to internal values are known, by applying successive basic equations in a step-by-step process.

It may be possible to use the differential equation, the boundary condition and the Taylor series to obtain a relation of the form

$$y_1 = Py_0+Q, \qquad (20)$$

where P, Q are constants. If this is used in place of (19), there is no longer a difference correction associated with the boundary condition. The additional work of deriving (20) may, however, outweigh this advantage.

SOLUTION OF THE ALGEBRAIC EQUATIONS

9. The solution of the equations $\mathbf{Ay} = \mathbf{d}$, when $\mathbf{A}$ has the form (6), can be effected in several ways. The method described in Chapter 1, §§ 10–12, which uses the decomposition $\mathbf{A} = \mathbf{LU}$ is very convenient; $\mathbf{L}$ and $\mathbf{U}$ each have non-zero elements only in the leading diagonal and one adjacent diagonal, and the equations determining the elements are correspondingly simple. In this particular case, moreover, the same equations are obtained by straightforward elimination without interchanges.

We eliminate the term in y_1 from the second equation

$$a_2y_1+b_2y_2+c_2y_3 = d_2, \qquad (21)$$

by subtracting the appropriate multiple a_2/b_1 of the first equation

$$b_1 y_1 + c_1 y_2 = d_1. \tag{22}$$

This leads to a new equation

$$\beta_2 y_2 + c_2 y_3 = \delta_2, \tag{23}$$

which in turn can be used to eliminate the term in y_2 from the third equation, and so on. We thus obtain a set of equations

$$\beta_r y_r + c_r y_{r+1} = \delta_r \quad (r = 1, 2, \ldots, n-2), \tag{24}$$

where β_r, δ_r are obtained from the recurrence relations

$$\beta_r = b_r - m_r c_{r-1}, \qquad \delta_r = d_r - m_r \delta_{r-1}, \qquad m_r = a_r/\beta_{r-1}, \tag{25}$$

with $\beta_1 = b_1, \delta_1 = d_1$. The final equation is

$$\beta_{n-1} y_{n-1} = \delta_{n-1}, \tag{26}$$

from which y_{n-1} may be obtained immediately. Then $y_{n-2}, y_{n-3}, \ldots, y_1$ may be obtained successively from the recurrence relation (24).

The process is well adapted to both desk-machine and automatic work. If the difference correction is now incorporated, the new values of y_r are obtained by use of (24) and the second of (25) only, together with the values of β_r, m_r already computed.

10. Iterative methods (see Chapter 4) can be used when the matrix is well-conditioned, and this is usually the case when the complementary function of the differential equation is of exponential type.

When the complementary function is of oscillatory character, however, iterative methods may converge slowly or even diverge. The method of elimination just described can still be used, though difficulties may occur if one or more of the β_r is very small. These difficulties have been discussed by Fox [76]. They do not arise, however, if interchanges are used in the elimination process; furthermore, the band structure of the equations is preserved. They are also avoided by using step-by-step methods, as described in §§ 11, 12. If β_{n-1} vanishes, $\mathbf{A}$ is singular and the equations $\mathbf{Ay} = \mathbf{d}$ have no solution.

USE OF STEP-BY-STEP METHODS

11. Boundary-value problems may also be solved by the step-by-step methods of Chapter 9. We compute a number of trial solutions which fulfil the boundary conditions at one end of the range, and combine them in such a way as to satisfy the conditions at the other end. In the case of a linear second-order differential equation, only two trial solutions have to be combined in this way.

In fact this method differs from that already discussed only in using a different procedure to solve the associated algebraic equations. In the case of equation (1) with the boundary values given, for example, the following sequence of operations is carried out.

(i) With an arbitrary y_1, we calculate by recurrence a particular solution $y^{(1)}$ satisfying the first $n-2$ equations of (4), with the difference correction neglected.

(ii) In a similar way we obtain a solution $y^{(2)}$ satisfying the first $n-2$ of the homogeneous equations corresponding to (4).

(iii) We now determine α such that $y^{(1)}+\alpha y^{(2)}$, which satisfies the first $n-2$ equations of (4) automatically, also satisfies the last equation.

(iv) Next, the difference correction $C(y^{(1)}+\alpha y^{(2)})$ is computed and inserted into the equations (4). The calculations of (i) and (iii) are then repeated, though since the homogeneous equations are unchanged, we use the same solution $y^{(2)}$. The difference correction should be calculated from the differences of $y^{(1)}+\alpha y^{(2)}$ rather than those of $y^{(1)}, y^{(2)}$ separately, since the latter may vary much more rapidly over the range of integration.

12. If the boundary condition at x_0 is of the more general form (16) then we modify our equations as indicated in § 8. The term in y_0 is transferred to the left-hand side of the first equation of (4) and an additional equation relating y_0 and y_1 is provided by (19) or (20). In either case the subsequent computation is carried out as in § 11, with the additional difference correction $C_1 y_0$ if (19) is used. Problems having the more general form of boundary condition at both ends can be solved in a similar way.

Several methods of this type have been discussed by Fox [76]. It should be noted that when the complementary functions are of exponential type there will often be a loss of significant figures when the trial solutions are combined; in such cases the method of § 9 is preferable.

METHOD OF CHEBYSHEV EXPANSION

13. No essential modification of the method of Chapter 9, §§ 23–25 is required in order to apply it to boundary-value problems associated with linear differential equations having polynomial coefficients. The boundary conditions give immediately two equations which must be satisfied by the coefficients in the expansion of the solution in Chebyshev series.

LINEAR DIFFERENTIAL EQUATIONS OF OTHER ORDERS

14. Similar methods can be used for equations of other orders and for simultaneous differential equations. In particular, a fourth-order equation may have two boundary conditions at each end-point, and the matrix **A** is a band matrix of width five. The algebraic equations can then be solved by matrix decomposition. In general, if **A** is a band matrix of width $2k+1$, where $k \geqslant 2$, the matrices **L** and **U** each have non-zero elements in the leading diagonal and k adjacent diagonals.

NON-LINEAR EQUATIONS

15. In the case of non-linear differential equations the algebraic equations resulting from the use of finite differences will also be non-linear. There is no established method for solving simultaneous non-linear algebraic equations. In many cases, however, the type of solution will be known from physical considerations, and iterative methods can be used. Whatever method of solution is used for the first approximation, successive corrections are usually obtained from linear equations, and the methods of this chapter are then applicable.

16. Problems of boundary-value type occurring, for example, in vibration theory lead to differential equations, usually linear and homogeneous, containing a parameter λ, and for which non-trivial solutions exist only for certain values of λ. The problem is to calculate one or more values of λ and the associated solutions.

The simplest example of this type is the equation

$$y''+\lambda y = 0, \tag{27}$$

with boundary conditions $y = 0$ at $x = 0$, $x = 1$. This problem has a known solution, non-trivial only if $\lambda = n^2\pi^2$ $(n = 1, 2, \ldots)$ when $y = \sin n\pi x$. The use of finite-difference equations leads to a matrix equation of the form

$$(\mathbf{A}-\lambda\mathbf{I})\,\mathbf{y} = 0, \tag{28}$$

and the problem reduces to that of calculating the latent roots and vectors of the matrix **A**, for which general methods have been discussed in Chapter 3. In the case of ordinary differential equations **A** is a band matrix and the smallest roots λ are the most important.

The use of relaxation for solving eigenvalue problems by finite-difference methods is described in [**76**] and [**77**].

17. Eigenvalue problems are also conveniently solved by the method of Chebyshev expansion if the coefficients in the differential equation are polynomials in x. The coefficients a_r in the expansion of y in series of Chebyshev polynomials are now linear functions of λ. The resulting infinite set of linear equations for the a_r and λ can be solved by an iterative method outlined in [**71**] or by direct methods. The latter approach yields an algebraic latent root problem, which resembles that of (28). We may expect, however, that for the same order matrix the Chebyshev method will yield more accurate eigenvalues λ because of the economy of its series representation.

11

HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

CLASSIFICATION OF PARTIAL DIFFERENTIAL EQUATIONS

1. We commence by indicating the fundamental classification of quasi-linear partial differential equations of the second order in two independent variables into *elliptic*, *parabolic* and *hyperbolic* types. The difference between these classes concerns the analytic character of their solutions, and the types of boundary conditions necessary to determine these solutions.

The general equation has the form

$$a\frac{\partial^2 u}{\partial x^2}+b\frac{\partial^2 u}{\partial x\,\partial y}+c\frac{\partial^2 u}{\partial y^2}=e, \tag{1}$$

where a, b, c and e are functions of u, $\partial u/\partial x, \partial u/\partial y, x$ and y, but not of the second derivatives. We shall adopt the standard notation in which

$$p=\frac{\partial u}{\partial x},\qquad q=\frac{\partial u}{\partial y},\qquad r=\frac{\partial^2 u}{\partial x^2},\qquad s=\frac{\partial^2 u}{\partial x\,\partial y},\qquad t=\frac{\partial^2 u}{\partial y^2}. \tag{2}$$

Then equation (1) becomes

$$ar+bs+ct=e. \tag{3}$$

2. Suppose we are given a curve in the (x,y) plane and values of u, p and q at all points on that curve. It is assumed that these satisfy the relation

$$du=\frac{\partial u}{\partial x}dx+\frac{\partial u}{\partial y}dy=p\,dx+q\,dy \tag{4}$$

along the curve, since otherwise p and q could not possibly be the derivatives of u. We might ask ourselves the question, "Do the values of u, p and q on this curve, together with the requirement that u satisfies the differential equation, enable us to determine r, s and t on the curve?" We must have

$$d\left(\frac{\partial u}{\partial x}\right)=\frac{\partial^2 u}{\partial x^2}dx+\frac{\partial^2 u}{\partial x\,\partial y}dy$$

or

$$dp=r\,dx+s\,dy, \tag{5}$$

and also

$$d\left(\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x\,\partial y}dx + \frac{\partial^2 u}{\partial y^2}dy$$

or

$$dq = s\,dx + t\,dy. \tag{6}$$

Equations (3), (5) and (6) form a set of three linear equations in three unknowns r, s and t. In general there exists one solution, so that unique values of the second derivatives are determined at each point of the curve. If, however, the determinant of the coefficients of r, s and t vanishes at any point, that is if

$$\begin{vmatrix} a & b & c \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} = 0, \tag{7}$$

then in general the equations (3), (5) and (6) have no solution for that point. For a solution to be possible we know from the theory of linear equations [**13**] that the rank of the matrix

$$\begin{bmatrix} e & a & b & c \\ dp & dx & dy & 0 \\ dq & 0 & dx & dy \end{bmatrix}$$

must be two. The rank will be two if any of the three relations

$$\begin{vmatrix} e & a & b \\ dp & dx & dy \\ dq & 0 & dx \end{vmatrix} = 0, \tag{8}$$

$$\begin{vmatrix} e & a & c \\ dp & dx & 0 \\ dq & 0 & dy \end{vmatrix} = 0, \tag{9}$$

or

$$\begin{vmatrix} e & b & c \\ dp & dy & 0 \\ dq & dx & dy \end{vmatrix} = 0 \tag{10}$$

is true, each of these relations being equivalent to the others if equation (7) holds. The latter may be written in the form

$$a(dy)^2 - b\,dy\,dx + c(dx)^2 = 0, \tag{11}$$

which is a quadratic equation in dy/dx. For a point (x, y) associated with given values of u, p and q, then according as b^2 is greater than, equal to, or less than $4ac$ there will be two directions for which (11) is satisfied, one direction or no possible direction respectively.

3. If we have a domain in the (x, y) plane in which u, p, q are defined and if $b^2 > 4ac$ at each point of that domain, then the differential equation is said to be ***hyperbolic*** in that domain. Similarly, if $b^2 = 4ac$ throughout the domain the equation is said to be ***parabolic***, and if $b^2 < 4ac$ the equation is said to be ***elliptic***. It is important to notice that the class to which the equation belongs may be dependent upon the solution. Thus the equation

$$\frac{\partial^2 u}{\partial x^2} + u\frac{\partial^2 u}{\partial y^2} = 0 \tag{12}$$

is elliptic in any domain over which the solution u is positive, and hyperbolic in any domain over which the solution is negative. If the equation is linear, that is if a, b and c in (1) are functions of x and y only, then for a given domain in the (x, y) plane the class of the differential equation is independent of the particular solution required and is determined in advance. Thus the equation

$$(1+x^2)\frac{\partial^2 u}{\partial x^2} + (1+y^2)\frac{\partial^2 u}{\partial y^2} = 0 \tag{13}$$

is always elliptic and the equation

$$\frac{\partial^2 u}{\partial x^2} + (1-x^2-y^2)\frac{\partial^2 u}{\partial y^2} = 0 \tag{14}$$

is elliptic inside the unit circle and hyperbolic outside.

DISCONTINUOUS SOLUTIONS

4. Returning to our original question we see that, when the relation (11) is satisfied along our curve, then in order for a solution to be possible we must insist that a further relation, given by any of the equivalent relations (8), (9) or (10), should also hold along our curve. If this is so, the theory of linear algebraic equations shows that there is an infinite number of solutions.

Let us suppose that we have chosen our initial curve and the values of u, p and q on it so that the relation (11) holds for all points on the curve. Such a curve, together with the values of u, p and q on it, is called a ***characteristic*** of the differential equation. (There is a lack of uniformity in the nomenclature used in the literature; some writers refer to the curve itself as a 'characteristic', and to the curve plus the values of u, p and q on it as a 'characteristic strip'.) For a solution of the differential equation to be possible, u, p and q must satisfy the further relation (8), say, and then we may choose one of r, s and t arbitrarily, the other two being determined uniquely. This means that we may have a curve C, lying in the domain of a solution u, such that the solution on both sides of the curve has the same values of u, p and q along the curve but different values of r, s and t. It is this important property which distinguishes hyperbolic partial differential equations from elliptic, since for elliptic equations such a situation is not possible. We may say that hyperbolic partial differential equations are characterized by the possession of real characteristics.

5. It is quite simple to construct a solution exhibiting the phenomenon just described. The two functions

$$u_1 = (x-y)^2+(x-y)+1$$

$$u_2 = 2(x-y)^2+(x-y)+1$$

satisfy $$u_1 = u_2 = 1 \quad \text{on} \quad x = y.$$

Also $$\partial u_1/\partial x = \partial u_2/\partial x = 1, \quad \partial u_1/\partial y = \partial u_2/\partial y = -1 \quad \text{on} \quad x = y.$$

The functions and their first derivatives therefore take the same values on the line $x = y$. They clearly also satisfy the differential equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2}$$

of which this line is a characteristic. The second derivatives, however, have quite different values on this line.

This phenomenon is of frequent occurrence in physical problems and is not merely of theoretical interest.

6. Another way of deriving the equations for the characteristic lines is to seek curves along which the partial differential equation reduces to an ordinary differential equation. If dx, dy, du, dp, dq denote differentials along any curve, C, then they must satisfy equations (5) and (6). Multiplying (5) by $a\,dy$, (6) by $c\,dx$ and adding the results, we have

$$a\,dy\,dp+c\,dx\,dq = (ar+ct)\,dx\,dy+as(dy)^2+cs(dx)^2. \tag{15}$$

Substitution from the differential equation (3) gives

$$\begin{aligned} a\,dy\,dp+c\,dx\,dq &= (-bs+e)\,dx\,dy+as(dy)^2+cs(dx)^2 \\ &= e\,dx\,dy+s\{a(dy)^2-b\,dx\,dy+c(dx)^2\}. \end{aligned} \tag{16}$$

If the curve is chosen so that

$$a(dy)^2-b\,dx\,dy+c\,(dx)^2 = 0, \tag{17}$$

then the term in s is eliminated and equation (16) reduces to

$$a\,dy\,dp+c\,dx\,dq = e\,dx\,dy, \tag{18}$$

which is an ordinary differential equation. Equations (17) and (18) are the *characteristic relations* (7) and (9) above.

7. In the rest of this chapter we shall discuss hyperbolic equations and some numerical methods for solving such equations. The treatment of parabolic and elliptic equations is discussed in Chapter 12.

SIMPLE EXAMPLE OF A HYPERBOLIC PARTIAL DIFFERENTIAL EQUATION

8. The simplest example of a hyperbolic partial differential equation is provided by a vibrating string. If u is the displacement, x is measured along the equilibrium position of the string and t is the time, then the differential equation is the simple wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}. \tag{19}$$

Putting $ct = y$ we obtain

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2}. \tag{20}$$

(The symbol t has been used in two senses, but since the independent variable t is now replaced by y this should cause no confusion.) The characteristic directions are given by the determinantal equation

$$\begin{vmatrix} 1 & 0 & -1 \\ dx & dy & 0 \\ 0 & dx & dy \end{vmatrix} = 0, \tag{21}$$

which reduces to the simple result $dy/dx = \pm 1$. The characteristic curves are therefore the straight lines $y = \pm x + \text{constant}$. For a solution to be possible we must also have

$$\begin{vmatrix} 0 & 1 & 0 \\ dp & dx & dy \\ dq & 0 & dx \end{vmatrix} = 0, \tag{22}$$

giving $dy/dx = dp/dq$, so that $dp/dq = \pm 1$ must be satisfied on the characteristics.

SOLUTION OF A SIMPLE DIFFERENTIAL EQUATION BY THE METHOD OF CHARACTERISTICS

9. So far, the notion of characteristics has been used to classify various types of differential equations. For hyperbolic equations the existence of real and distinct characteristics leads to the most satisfactory known method of numerical solution. The relation (8) or its equivalents must be satisfied along characteristic lines. From the equations (7) and (8) we can construct the characteristics and the solution to the differential equation on these lines.

10. Before describing the general method it will be helpful to consider the simple case of the vibrating string, satisfying the differential equation (19). Suppose that we have a string fixed at both ends, $x = 0$ and $x = 1$ say, and that initially both its form u and velocity $\partial u/\partial t$ are specified. The 'boundary conditions' for equation (20) are then

$$u = 0 \quad \text{at} \quad x = 0 \quad \text{and} \quad 1, \quad 0 \leqslant y \leqslant \infty, \tag{23}$$

and hence

$$q = \partial u/\partial y = 0 \quad \text{at} \quad x = 0 \quad \text{and} \quad 1, \quad 0 \leqslant y \leqslant \infty. \tag{24}$$

The 'initial conditions' specify u, and hence $p = \partial u/\partial x$, and q at $y = 0$ for $0 \leqslant x \leqslant 1$.

Suppose we wish to know the displacement and velocity of the points of the string at some later time given by $y = k$. It has already been shown that the characteristics are the lines $x \pm y = \text{constant}$, and that the relations $p \pm q = \text{constant}$ must be satisfied on them. If in Figure 1 we take any point P on $y = k$ we can draw two characteristics through it. The line PP_3, which satisfies $x + y = \text{constant}$, meets $x = 1$ at P_3. We can draw through P_3 the characteristic P_3P_1 of type $x-y =$ constant, meeting the line AB, or $y = 0$, at P_1. (In general, for a large value of k, we shall hit the boundaries $x = 0$ and $x = 1$ a number of times before we reach the line AB.) Similarly, if we draw the other characteristic through P we ultimately reach AB at P_2.

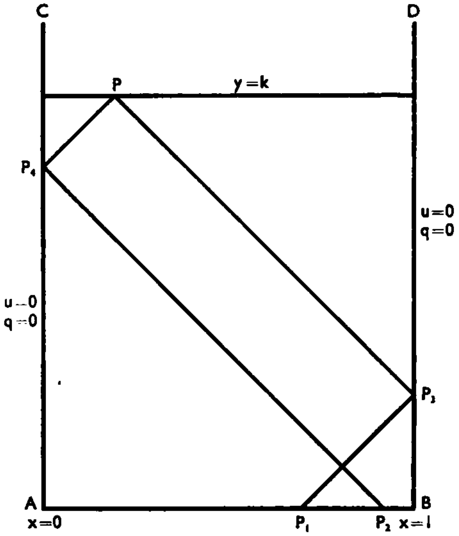


Figure 1

11. At P_1 the values of p and q, say p_1 and q_1, are given by the initial conditions. Then on P_1P_3 we have

$$p-q = \text{constant} = p_1-q_1. \tag{25}$$

At the point P_3 on BD we know $q = 0$, from the boundary conditions (24), and therefore

$$p = p_1-q_1. \tag{26}$$

On the line PP_3 we have

$$p+q = \text{constant} = p_1-q_1 \tag{27}$$

in virtue of (26) and (24). Similarly, starting from P_2 at which p and q have known values p_2, q_2, say, we have, on P_2P_4,

$$p+q = \text{constant} = p_2+q_2. \tag{28}$$

At the point P_4 on AC we know $q = 0$, so that, at P_4,

$$p = p_2+q_2. \tag{29}$$

On P_4P we have
$$p-q = \text{constant} = p_2+q_2 \tag{30}$$

in virtue of (29) and (24). Equations (27) and (30) are both satisfied at P, so that

$$\left.\begin{aligned} p &= \tfrac{1}{2}(p_1-q_1+p_2+q_2), \\ q &= \tfrac{1}{2}(p_1-q_1-p_2-q_2). \end{aligned}\right\} \tag{31}$$

We may find p and q at all points on $y = k$ in a similar manner. The quantity q gives us the velocity, and by integrating p with respect to x, starting at $x = 0$ where $u = 0$, we obtain u at all points. At $x = 1$ we should obtain the given boundary value $u = 0$.

12. The solution of this example is very simple because the characteristics in the (x, y) plane are determined independently of the solution and are straight lines. The general case has neither of these simplifications.

THE SOLUTION BY CHARACTERISTICS IN THE GENERAL CASE

13. Suppose we are given the values of u, p, q on a curve AB in Figure 2 which is *not* a characteristic curve. This restriction is important in view of the considerations of § 4. If we take two points P and Q on AB, then there are two characteristics through P and and two through Q. The directions of the two characteristics are given by equation (11), whose solutions are

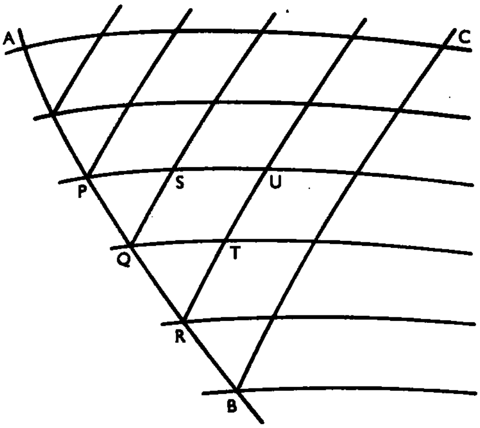


Figure 2

$$\frac{dy}{dx} = \frac{b \pm (b^2 - 4ac)^{\frac{1}{2}}}{2a}. \tag{32}$$

These may be represented by

$$\frac{dy}{dx} = f \quad \text{and} \quad \frac{dy}{dx} = g. \tag{33}$$

The relation (9) which must be satisfied on a characteristic gives

$$e\,dx\,dy - a\,dp\,dy - c\,dq\,dx = 0. \tag{34}$$

The characteristic PS of the f type meets the characteristic QS of the g type at S. If P is close to Q, then as a first approximation we may regard PS as a straight line of slope f_P. We have therefore as a first approximation the equation

$$y_S - y_P = f_P(x_S - x_P). \tag{35}$$

Similarly, regarding QS as straight, we have

$$y_S - y_Q = g_Q(x_S - x_Q). \tag{36}$$

Equations (35) and (36) are a pair of linear equations in the two unknowns y_S and x_S, which can therefore be determined. Using these initial approximations to x_S and y_S we may approximate to (34) by the equations

$$\left.\begin{array}{l} e_P(y_S - y_P) - a_P(p_S - p_P)f_P - c_P(q_S - q_P) = 0 \text{ (along } PS\text{),} \\ \text{and by} \\ e_Q(y_S - y_Q) - a_Q(p_S - p_Q)\, g_Q - c_Q(q_S - q_Q) = 0 \text{ (along } QS\text{).} \end{array}\right\} \tag{37}$$

This is a pair of linear equations in p_S and q_S from which first approximations to these values may be found. We may then obtain u_S from the relation

$$u_S = u_P + dx\left(\overline{\frac{\partial u}{\partial x}}\right) + dy\left(\overline{\frac{\partial u}{\partial y}}\right), \tag{38}$$

where $\overline{\dfrac{\partial u}{\partial x}}, \overline{\dfrac{\partial u}{\partial y}}$ are mean values along PS. This gives

$$u_S = u_P + (x_S - x_P)\tfrac{1}{2}(p_S + p_P) + (y_S - y_P)\tfrac{1}{2}(q_S + q_P). \tag{39}$$

From the approximations to u_S, p_S and q_S so obtained we can approximate to f_S and g_S from (32) and (33).

14. We may now obtain improved values for x_S, y_S, p_S, q_S, u_S in the following way. As improved versions of (35) and (36) we have

$$\left.\begin{array}{l} y_S - y_P = \tfrac{1}{2}(f_S + f_P)(x_S - x_P), \\ y_S - y_Q = \tfrac{1}{2}(g_S + g_Q)(x_S - x_Q), \end{array}\right\} \tag{40}$$

from which we calculate more accurate values of x_S and y_S. Similarly the improved version of the first of (37) is

$$\tfrac{1}{2}(e_P + e_S)(y_S - y_P) - \tfrac{1}{2}(a_P + a_S)(p_S - p_P)\tfrac{1}{2}(f_P + f_S) - \tfrac{1}{2}(c_P + c_S)(q_S - q_P) = 0, \tag{41}$$

with a similar improvement for the second of (37). From these two equations we obtain more accurate values of p_S and q_S and finally of u_S as before. The process should be repeated until two successive approximations to x_S, y_S, p_S, q_S, u_S agree to the accuracy to which we are working. Normally we would take the point P so close to Q that one improvement gives the desired accuracy. When this accuracy has been attained at S and T we can proceed a step further, to the point U in Figure 2, and so on.

15. Our method, then, is to take a number of points on our initial curve at convenient distances apart and to obtain the values of x, y, u, p, q at all points of the mesh shown in Figure 2. It will be seen that the curve AB and the values of u, p and q on it only determine the values of u, p and q in the curvilinear triangle ABC bounded by the two characteristics AC and BC. Points outside this triangle are influenced by values on the continuation of the curve AB. The values at the points in ABC are completely independent of the values of u, p and q at points on the initial curve beyond A and B.

16. Apart from the ordinary linear wave equations of which the vibrating string is a simple example, perhaps the most important example of a hyperbolic partial differential equation is provided by steady supersonic compressible flow in two dimensions. If we have isentropic potential flow, then there is a velocity potential u, and the pressure P is a function, $P(\rho)$, of the density ρ only. The quantity $dP/d\rho$ is equal to a^2, where $a = a(\rho)$ is the velocity of sound. The equation satisfied by the potential u is

$$(a^2-p^2)\frac{\partial^2 u}{\partial x^2}-2pq\frac{\partial^2 u}{\partial x\,\partial y}+(a^2-q^2)\frac{\partial^2 u}{\partial y^2}=0, \tag{42}$$

where

$$\left.\begin{aligned} p &= \frac{\partial u}{\partial x} = -(\text{velocity in } x \text{ direction}),\\ q &= \frac{\partial u}{\partial y} = -(\text{velocity in } y \text{ direction}).\end{aligned}\right\} \tag{43}$$

There is a further relation (Bernoulli's equation) given by

$$\int\frac{dP}{\rho} = \text{constant}-\tfrac{1}{2}(p^2+q^2). \tag{44}$$

The left-hand side of (44) is a function of ρ, so that (44) gives ρ in terms of p and q, and since a is a function of ρ it is also a function of p and q.

17. The characteristic directions (Mach lines) are given by

$$(a^2-p^2)(dy)^2+2pq\,dx\,dy+(a^2-q^2)(dx)^2 = 0. \tag{45}$$

These are real if $$p^2q^2>(a^2-p^2)(a^2-q^2),$$

that is, if $$p^2+q^2>a^2. \tag{46}$$

Now p^2+q^2 is the square of the velocity, so that we see from (46) that the equation is hyperbolic if the motion is supersonic. The relation to be satisfied on the characteristics is

$$\begin{vmatrix} 0 & (a^2-p^2) & (a^2-q^2)\\ dp & dx & 0\\ dq & 0 & dy\end{vmatrix} = 0, \tag{47}$$

giving $$-(a^2-p^2)\,dp\,dy = (a^2-q^2)\,dq\,dx,$$

or $$-\frac{dy}{dx} = \frac{(a^2-q^2)}{(a^2-p^2)}\frac{dq}{dp}. \tag{48}$$

From (45) we then find

$$(a^2-q^2)(dq/dp)^2-2pq(dq/dp)+(a^2-p^2) = 0, \tag{49}$$

for the relations to be satisfied on the characteristics. Since, however, a is a function of p and q, (49) can be integrated (in general only numerically), and hence the relation between p and q along a characteristic is independent of the curve in the (x, y) plane. This is always true of the homogeneous quasi-linear second-order equation (3) in which the coefficients of r, s and t are functions of p and q only. It is clear from (45) and (49) that if the directions of the characteristics are given by $dy/dx = f$, $dy/dx = g$, then along them we have $dq/dp = -1/g$ and $dq/dp = -1/f$ respectively.

SIMULTANEOUS PARTIAL DIFFERENTIAL EQUATIONS

18. The method of characteristics may be applied to a system of simultaneous first-order differential equations. An example will make this clear. Suppose we consider the non-steady one-dimensional motion of compressible fluid in which $P = P(\rho)$, so that $dP/d\rho = a^2$, the square of the velocity of sound. The equation of motion is given by

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+\frac{a^2}{\rho}\frac{\partial \rho}{\partial x} = 0, \tag{50}$$

and the equation of continuity is

$$\rho\frac{\partial u}{\partial x}+\frac{\partial \rho}{\partial t}+u\frac{\partial \rho}{\partial x} = 0. \tag{51}$$

Again we may pose the question: if we are given the values of u and ρ on a curve in the (x,t) plane, do the differential equations determine $\frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}, \frac{\partial \rho}{\partial t}$ and $\frac{\partial \rho}{\partial x}$ on this line? The equations

$$\frac{\partial u}{\partial t}dt+\frac{\partial u}{\partial x}dx = du, \tag{52}$$

$$\frac{\partial \rho}{\partial t}dt+\frac{\partial \rho}{\partial x}dx = d\rho, \tag{53}$$

must be satisfied, and together with (50) and (51) they provide four equations in the four unknowns $\frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}, \frac{\partial \rho}{\partial t}$ and $\frac{\partial \rho}{\partial x}$, for which in general a unique solution will be obtained. If, however, the relation

$$\begin{vmatrix} 1 & u & 0 & \frac{a^2}{\rho} \\ 0 & \rho & 1 & u \\ dt & dx & 0 & 0 \\ 0 & 0 & dt & dx \end{vmatrix} = 0 \tag{54}$$

is satisfied, the equations will not have a solution unless

$$\begin{vmatrix} 1 & u & 0 & 0 \\ 0 & \rho & 1 & 0 \\ dt & dx & 0 & du \\ 0 & 0 & dt & d\rho \end{vmatrix} = 0, \tag{55}$$

or any of the three equivalent relations is satisfied. Equation (54) reduces to $dx/dt = u \pm a$. This means that for all one-dimensional unsteady flow problems there are two real characteristic directions, it being natural to use the same nomenclature here as in the quasi-linear second-order case. The relation (55) reduces to

$$\frac{du}{d\rho} = \mp \frac{a}{\rho}, \tag{56}$$

for $dx/dt = u \pm a$. Equation (56) gives

$$u = \mp \int \frac{a}{\rho} d\rho, \tag{57}$$

and, since a is a function of ρ, u is also a function of ρ. The pair of differential equations (50) and (51) may be solved by the method of characteristics in exactly the same manner as in the case of a second-order partial differential equation already described.

COMPARISON WITH ELLIPTIC EQUATIONS

19. The above examples manifest a feature which distinguishes the formulation of problems involving hyperbolic and parabolic equations from those involving elliptic equations, namely that the boundary conditions are commonly specified on an open boundary. For elliptic equations it is usual to have closed boundaries. A typical problem in the elliptic field is: "Given u on a closed curve, to find u inside that curve to satisfy the boundary condition and the differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

inside the curve". An interesting feature of this problem is that even if the values of u on the boundary are discontinuous at a number of points, the solution u has derivatives of all orders in both x and y inside the curve. This is in striking contrast to the hyperbolic case, where boundary conditions with discontinuities in the derivatives give rise to solutions with discontinuities in the derivatives. The well-behaved nature of the solution in the elliptic case has the result that finite-difference techniques are far less likely to lead to difficulties and it is usually quite safe to use a rectangular mesh of points. For hyperbolic equations the possibility of having discontinuities in the second derivative across the characteristics makes the use of a rectangular mesh rather hazardous, and it is better to use the characteristic mesh in spite of the fact that it provides values of u, p and q at the rather inconveniently placed points of intersection; see also [81].

12

PARABOLIC AND ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

BOUNDARY CONDITIONS

1. In Chapter 11 the three types of partial differential equations, hyperbolic, parabolic and elliptic, were distinguished by reference to their characteristics. For hyperbolic equations the characteristics were real curves, and the chosen numerical method of solution involved a step-by-step process carried out along these curves; this method was possible and convenient because the supplementary conditions were of *initial-value* type and the boundary was *open.*

With elliptic equations the conditions are of *boundary-value* type, generally given at all points of a *closed* boundary. There are corresponding distinctions in the methods of solution, step-by-step methods being replaced by the simultaneous solution of the relevant finite-difference equations.

Parabolic equations come somewhere between these two extremes. The boundary is open, but usually only in one direction, and the best methods of solution are combinations of boundary-value and initial-value techniques.

PARABOLIC EQUATIONS

2. Physical problems leading to parabolic equations are those of heat conduction and diffusion. The simplest equation of this kind is given by

$$\frac{\partial^2 f}{\partial x^2}=\frac{\partial f}{\partial t}, \tag{1}$$

with conditions specifying f on three sides of a rectangle, $t = 0$, $x = -1$, $x = +1$, as shown in Figure 1. The boundary is open in the positive t-direction.

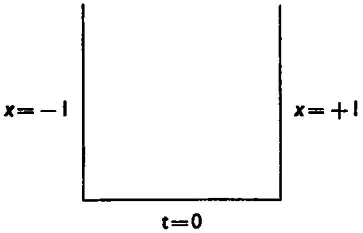


Figure 1

It is worth noticing in passing that the equation as presented is dimensionless. The original equation might be given as

$$k\frac{\partial^2 f}{\partial x^2}=\frac{\partial f}{\partial t}, \tag{2}$$

with the boundaries at $x = -l$ and $+l$. In the heat conduction equation f is temperature, x is length, t is time and k, called the *diffusivity* or *thermometric conductivity*, has the dimensions of $\mathrm{L}^2\mathrm{T}^{-1}$. If we introduce the new independent variables $X = (1/l)x, T = (k/l^2)t$, the equation reduces to (1) with x and t replaced by X and T, and the boundaries are at $T = 0$, $X = \pm 1$. Such 'non-dimensional' treatment is often of considerable value in numerical work.

Analytical solutions given, for example, by Carslaw and Jaeger [**90**], can sometimes be obtained to equations like (1) with various boundary conditions. As with ordinary differential equations, however, these are often rather complicated expressions, whose evaluation is not trivial: small changes in the equations or boundary conditions, moreover, may prohibit the production of such a solution. Again, therefore, numerical methods of solution, based on the use of finite differences, have been developed. We shall use the simple equation (1) to illustrate methods that are applicable to parabolic equations of much more general form.

REDUCTION TO ORDINARY DIFFERENTIAL EQUATIONS

3. We first replace the second derivative in the x-direction by finite differences. We divide the range -1 to 1 into equal intervals δx with pivotal points $x_0, x_1, \ldots, x_n$, the first and last points lying on the boundaries, and we denote by $f_r(t)$ the function f evaluated as a function of t for the constant value x_r of x. We can then replace (1) by the equations

$$(\delta x)^2\frac{df_r}{dt} = (f_{r-1}-2f_r+f_{r+1})+C_x f_r \quad (r = 1, 2, \ldots, n-1). \tag{3}$$

Here $C_x f_r$ is the difference correction

$$C_x = -\tfrac{1}{12}\delta_x^4+\tfrac{1}{90}\delta_x^6-\ldots, \tag{4}$$

the suffix x referring to differences in the x-direction.

Equations (3) represent a set of ordinary differential equations and, if the difference correction is neglected, they can be written as

$$\left.\begin{array}{llllll}
p(df_1/dt)+2f_1- f_2 & & & & &= f_0,\\
p(df_2/dt)- f_1+2f_2- f_3 & & & & &= 0,\\
p(df_3/dt) & - f_2+2f_3 & - f_4 & & &= 0,\\
\multicolumn{6}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots}\\
p(df_{n-2}/dt) & & & - f_{n-3}+2f_{n-2}- f_{n-1} & &= 0,\\
p(df_{n-1}/dt) & & & & - f_{n-2}+2f_{n-1} &= f_n,
\end{array}\right\} \tag{5}$$

where $p = (\delta x)^2$. There are $n-1$ equations in $n-1$ unknown functions, and the extra conditions provide known values of all the f_r at $t = 0$; f_0 and f_n are known for all t.

Equations in this form have been used to solve problems in heat conduction on the differential analyser [**92**]. They can be solved on desk machines or automatic machines, by the methods of Chapter 9, the application of some of which we describe below.

4. If the conditions on $x = \pm 1$ do not specify the function, but involve some law of cooling represented by

$$\frac{\partial f}{\partial x} + kf = g, \tag{6}$$

say, with k and g known functions of t, we use the method given in Chapter 10, § 8 for the similar problem of ordinary differential equations involving derivative conditions at the boundaries. The set (5) will then have two extra equations involving df_0/dt and df_n/dt.

The centre line $x = 0$ is often a line of symmetry, and we can take advantage of this fact to halve the number of equations in (5).

5. For an equation in cylindrical coordinates, given by

$$\frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} = \frac{\partial f}{\partial t}, \tag{7}$$

we replace both the second and first r-derivatives by their simplest finite-difference approximations, and can again produce equations similar to (5). The line $r = 0$ is one of symmetry, and for the equation on this line we replace (7) by

$$p\frac{df_0}{dt} = 4(f_1 - f_0), \tag{8}$$

where $p = (\delta r)^2$ and the suffixes 0 and 1 here refer to the centre line and the adjacent line in the field of integration.

USE OF THE RUNGE–KUTTA METHOD

6. The Runge–Kutta process, which is well suited to automatic work, is immediately applicable to the set of equations (5). Unfortunately, however, though the truncation error in the fourth-order process can be made negligible without prohibitive reduction of the interval, the stability requirement severely restricts the size of interval δt that may be used.

The homogeneous equations obtained from (5) by neglecting the right-hand sides, have solutions of the form $f_r = a_r e^{-\lambda t}$ where λ and a_r $(r = 1, 2, \ldots, n-1)$ are constants satisfying

$$\left.\begin{aligned} (2-\lambda p)a_1 - a_2 &= 0,\\ -a_1 + (2-\lambda p)a_2 - a_3 &= 0,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots&\\ -a_{n-2} + (2-\lambda p)a_{n-1} &= 0. \end{aligned}\right\} \tag{9}$$

These equations in turn are satisfied by $a_r = \sin r\theta$, provided that $2 - \lambda p = 2\cos\theta$ and $\sin n\theta = 0$, that is

$$\lambda = \frac{4}{p}\sin^2\frac{r\pi}{2n} = \frac{4}{(\delta x)^2}\sin^2\frac{r\pi}{2n} \quad (r = 1, 2, \ldots, n-1). \tag{10}$$

The largest value of λ is approximately $4/(\delta x)^2$. Since the fourth-order Runge–Kutta process is stable for an interval h, in this case δt, such that $\lambda h < 2{\cdot}8$ (see Chapter 9, § 29), we see that this leads to the restriction

$$\frac{\delta t}{(\delta x)^2} < 0{\cdot}7. \tag{11}$$

If the partial differential equation is not of the simple form (1) but contains rapidly varying functions, so that a fairly small interval δx has to be used, the restriction (11) on the interval δt may be prohibitive. For this reason it may be desirable, even when using an automatic computer, to keep the interval δx as large as possible and incorporate the difference correction $C_x f$.

AN EXPLICIT METHOD

7. This method replaces the time-derivative in equations (3) by its simplest finite-difference approximation

$$\delta t\left(\frac{\partial f_r}{\partial t}\right)_{t_0} = f_{r,t_0+\delta t} - f_{r,t_0}. \tag{12}$$

Then neglecting the difference corrections, we have

$$f_{r,t_0+\delta t} = f_{r,t_0} + \frac{\delta t}{(\delta x)^2}(f_{r-1} - 2f_r + f_{r+1})_{t_0} \quad (r = 1, 2, \ldots, n-1). \tag{13}$$

The values of f_r at time $t_0+\delta t$ are thus obtained directly from those at time t_0, and for this reason the method is said to be *explicit*.

Though the truncation error in (12) is much greater than that involved in the Runge–Kutta process, it is usually the stability requirement that governs the permissible size of δt. An analysis similar to that of the previous section shows that this method is stable provided that

$$\frac{\delta t}{(\delta x)^2} < \tfrac{1}{2}. \tag{14}$$

This restriction is only slightly more stringent than that of (11) for the Runge–Kutta process and, of course, the present method involves less computation. Again it would be possible to include the difference corrections, but the choice of an interval δt for which $C_t f$ must be included is only to be recommended for desk-machine work.

RICHARDSON'S METHOD

8. If (12) is replaced by the better approximation

$$2\delta t\left(\frac{\partial f_r}{\partial t}\right)_{t_0} = f_{r,t_0+\delta t} - f_{r,t_0-\delta t}, \tag{15}$$

we obtain, in place of (13),

$$f_{r,t_0+\delta t} = f_{r,t_0-\delta t} + \frac{2\delta t}{(\delta x)^2}(f_{r-1} - 2f_r + f_{r+1})_{t_0}. \tag{16}$$

This method, however, is unstable for all values of $\delta t/(\delta x)^2$, and so should not be used.

9. This process is obtained by applying to the set of equations (3) the method outlined in Chapter 9, § 16 for the solution of first-order ordinary differential equations. Thus, combining (3) with the relations

$$\left.\begin{aligned} f_{t_0+\delta t}-f_{t_0} &= \tfrac{1}{2}\delta t\left\{\left(\frac{\partial f}{\partial t}\right)_{t_0+\delta t}+\left(\frac{\partial f}{\partial t}\right)_{t_0}\right\}+(C_t f)_{t_0+\frac{1}{2}\delta t}, \\ \text{where}\qquad C_t &= -\tfrac{1}{12}\delta_t^3+\tfrac{1}{120}\delta_t^5-\ldots, \end{aligned}\right\} \tag{17}$$

we derive

$$\begin{aligned} &\{f_{r-1}-2(1+s)f_r+f_{r+1}+C_x f_r\}_{t_0+\delta t} \\ &\quad = -\{f_{r-1}-2(1-s)f_r+f_{r+1}+C_x f_r\}_{t_0}-2s(C_t f_r)_{t_0+\frac{1}{2}\delta t}\quad (r=1,2,\ldots,n-1), \end{aligned} \tag{18}$$

where $s = (\delta x)^2/\delta t$.

To obtain the values of f at the time $t_0+\delta t$ we then have to solve a set of simultaneous equations; this is accordingly an *implicit* method. If the difference correction $C_x f$ is neglected the matrix of coefficients of these equations is a band matrix of width three and the equations can be solved very simply by the method given in Chapter 10, § 9.

It may be noted that $(C_x f_r)_{t_0}$ will often be a good approximation to $(C_x f_r)_{t_0+\delta t}$, and may be used to replace it in (18); in this way the difference correction $C_x f$ is effectively incorporated in a single integration run. Determination, or estimation, of the difference correction $C_t f$ is less convenient, and on an automatic computer it is customary to use an interval δt sufficiently small for $C_t f$ to be assumed negligible. The validity of this assumption can be checked by a second run using a different interval δt.

Though this method involves a good deal more work at each step it has the great advantage, shared with the corresponding method of Chapter 9, of being stable for all values of the intervals δx and δt. The number of time steps needed is often considerably less than with other methods and in consequence the Crank–Nicolson method is usually to be recommended.

NON-LINEAR EQUATIONS

10. If the differential equation, while retaining its parabolic form, is non-linear or has non-linear boundary conditions, then the methods of §§ 6 and 7 are immediately applicable. The Crank–Nicolson method, however, involves the solution of a set of simultaneous non-linear algebraic equations. This may be accomplished by extrapolating from the known values of f_r for $t_0, t_0-\delta t, \ldots$ to obtain an estimate of f_r for $t_0+\delta t$, and then applying Newton's rule (Chapter 6, § 9) to obtain an accurate solution of the equations; one or two applications normally suffice. Again, despite the added complexity of the calculations, this method is often much faster than those subject to restrictive stability limitations.

SINGULARITIES

11. All methods may have difficulties when there are singularities in the boundary conditions, for example with 'quenching' problems in which a high temperature at $t = 0$ is reduced instantaneously to zero on the

boundaries $x = \pm 1$. Near the corner points $t = 0, x = \pm 1$, the finite-difference equations cease to be meaningful, and the step-by-step process is either replaced in the early stages by an analytical solution calculable for small t, or is carried out following a transformation of both independent variables which removes the singularity.

When the singularity is at $x = 0, t = 0$, the transformation usually employed is

$$X = \frac{x}{2\sqrt{t}}, \qquad T = \sqrt{t}. \tag{19}$$

Equation (1) becomes

$$\frac{\partial^2 f}{\partial X^2} + 2X\frac{\partial f}{\partial X} = 2T\frac{\partial f}{\partial T}, \tag{20}$$

and the point $x = 0, t = 0$ is 'stretched out' into the X-axis.

In cases when the transient phenomena are not of interest it is sometimes possible to ignore the meaningless nature of the finite-difference representation in the neighbourhood of the singularity, because the errors so introduced do not persist; see for example [**96**].

ELLIPTIC EQUATIONS

12. Elliptic differential equations are of pure boundary-value type. The simplest equation of this kind is the famous equation of Laplace, given in two dimensions by

$$\nabla^2 f = \partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 = 0. \tag{21}$$

A more general equation is that of Poisson, given by

$$\nabla^2 f = g(x, y), \tag{22}$$

in which the right-hand side is a known function. With such equations are associated boundary conditions at all points of a closed boundary, specifying values of the function, or its normal derivative, or a combination of these quantities.

The general method of solution is to divide the region into a square or rectangular grid of pivotal points, replace the differential equation by finite-difference equations, and solve the resulting set of algebraic equations.

FINITE-DIFFERENCE REPRESENTATIONS

13. The simplest problem of this type is given by equation (21), with f specified at points on a rectangular boundary. If we consider a particular point (x_0, y_0) of the grid dividing up the rectangle, we can use the central-difference formula

$$(\delta x)^2\, \partial^2 f_0/\partial x^2 = (\delta_x^2 - \tfrac{1}{12}\delta_x^4 + \tfrac{1}{90}\delta_x^6 - \dots)f_0, \tag{23}$$

the suffix x denoting, as before, 'differencing in the x-direction'. A similar formula holds for the second derivative in the y-direction. As in the corresponding solution of ordinary differential equations of boundary-value type we replace the leading terms in the derivative formulae by their expressions in terms of pivotal values, given by

$$\delta_x^2 f_0 = (f_1 - 2f_0 + f_{-1})_x, \quad \delta_y^2 f_0 = (f_1 - 2f_0 + f_{-1})_y, \tag{24}$$

where the values $f_{1,x}$ and $f_{-1,x}$ belong to the pivotal points $x_0+\delta x$, $x_0-\delta x$, with similar meanings for $f_{1,y}$ and $f_{-1,y}$. We can then replace the differential equation by the difference equation

$$(f_1-2f_0+f_{-1})_x+s^2(f_1-2f_0+f_{-1})_y+Cf_0 = 0, \tag{25}$$

where $s = \delta x/\delta y$ and Cf_0 is the difference correction. In the most common case, in which $\delta x = \delta y = h$, equation (25) becomes

$$\left.\begin{aligned}(f_1+f_{-1})_x+(f_1+f_{-1})_y-4f_0+Cf_0 = 0,\\ C = (-\tfrac{1}{12}\delta^4+\tfrac{1}{90}\delta^6-\ldots)_x+(-\tfrac{1}{12}\delta^4+\tfrac{1}{90}\delta^6-\ldots)_y.\end{aligned}\right\} \tag{26}$$

The differential equation (22) leads to the same finite-difference equation, with the addition of the term $h^2g(x_0, y_0)$ on the right of the first of (26).

14. If the boundary values are known, an equation of type (26) is to be satisfied at every internal point of the mesh, and the number of unknowns is the same as the number of equations. This may be fairly large, but each equation contains at most five unknowns and the set is fairly well conditioned. They can usually be solved quite conveniently on an automatic computer either by direct methods or by the iterative methods described in Chapter 4, and on desk machines by direct methods or relaxation.

15. The direct methods are exemplified by Poisson's equation with a rectangular boundary, which leads to algebraic equations which can be represented by the matrix equation

$$\mathbf{Af} = \mathbf{b}, \tag{27}$$

in which $\mathbf{A}$ has the following special form in partitioned-matrix notation:

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \ldots \\ \mathbf{I} & \mathbf{B} & \mathbf{I} & \mathbf{O} & \mathbf{O} & \ldots \\ \mathbf{O} & \mathbf{I} & \mathbf{B} & \mathbf{I} & \mathbf{O} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}. \tag{28}$$

$\mathbf{I}$ is a unit matrix, and $\mathbf{B}$ a band matrix given by

$$\mathbf{B} = \begin{bmatrix} -4 & 1 & 0 & 0 & 0 & \ldots \\ 1 & -4 & 1 & 0 & 0 & \ldots \\ 0 & 1 & -4 & 1 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}. \tag{29}$$

The order of $\mathbf{A}$ is the total number of internal mesh points, while the order of $\mathbf{I}$ and $\mathbf{B}$ is the number of internal mesh points in one direction.

Descriptions of methods of solving (27) in these circumstances have been given by Karlqvist [**110**] and Cornock [**111**], with extensions to fourth-order equations of biharmonic type (see § 19).

16. As with ordinary differential equations our general procedure is to choose a reasonably large interval, to keep the number of algebraic equations as small as possible. In a first approximation the difference

correction is neglected and a solution $f^{(1)}$ produced. A few values can be found at external points by a step-by-step process, enough to enable the significant central differences of $f^{(1)}$ to be obtained at all internal pivotal points, and $Cf^{(1)}$ can then be calculated at all these points. Insertion of $Cf^{(1)}$ in the first of (26), with all constant terms suppressed, then provides a correction $\delta f^{(1)}$, and the process can be repeated if necessary.

When an automatic computer is used, it is customary to choose an interval for which the difference correction is expected to prove negligible. If the differences of the solution so obtained show that this is not so, then, even if storage limitations prevent the choice of a smaller interval, the difference corrections can be incorporated fairly readily. In such a case direct methods of solution of the algebraic equations have the advantage that the triangular decomposition of **A** does not have to be repeated.

BOUNDARY CONDITIONS INVOLVING A DERIVATIVE

17. If the boundary condition involves a derivative on the rectangular boundary, the procedure is very similar to that for ordinary differential equations. The finite-difference equation (26) is now used at a boundary point also, and the external value so introduced is eliminated by use of the finite-difference form of the boundary condition, involving a difference correction of new type at a boundary point.

A new problem is presented if the boundary is curved. Few boundary points are now mesh points, and the simple formula (26) cannot be used at points like O in Figure 2, since point 1 is outside the boundary. The boundary value (×) is known, so that using an interpolation formula we can express the value f_1 in terms of the boundary value and values at internal points, thereby eliminating f_1 from the basic finite-difference equation.

If the normal derivative is involved in the condition at a curved boundary the problem is much more difficult, since the direction of the normal does not coincide with that of either set of mesh lines.

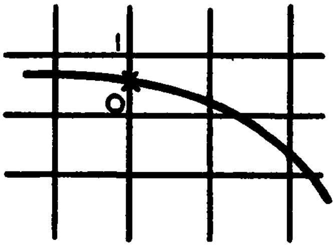


Figure 2

MORE ACCURATE FINITE-DIFFERENCE REPRESENTATION

18. For equations involving the Laplace operator we can construct more accurate but more complicated finite-difference expressions for the quantity $\nabla^2 f$. In the notation of Figure 3 we find, neglecting sixth and

higher differences, the formula

$$6h^2\nabla^2 f_0+\tfrac{1}{2}h^4\nabla^4 f_0 = 4(f_1+f_2+f_3+f_4)+(f_5+f_6+f_7+f_8)-20f_0. \quad (30)$$

Here both terms on the left-hand side are known, $\nabla^4 f_0$ being replaced by zero for Laplace's equation and by $\nabla^2 g_0$ for Poisson's equation (22). Other formulae and their method of derivation have been given by Bickley [**101**].

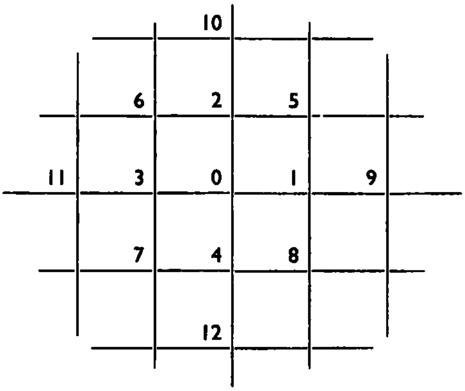


Figure 3

For more general equations such as

$$g_1\frac{\partial^2 f}{\partial x^2}+g_2\frac{\partial^2 f}{\partial y^2}+g_3\frac{\partial f}{\partial x}+g_4\frac{\partial f}{\partial y}+g_5 f = g_6, \quad (31)$$

where the g_r are known functions of x and y, the replacement of all the derivatives by their central-difference equivalents and the expression of their first terms by pivotal values will lead to a five-term formula like (26), though all the coefficients may be different. More accurate simple formulae like (30) are unlikely to exist in the general case.

OTHER PROBLEMS OF ELLIPTIC TYPE

19. The fourth-order equation

$$\nabla^4 f \equiv \frac{\partial^4 f}{\partial x^4}+2\frac{\partial^4 f}{\partial x^2\partial y^2}+\frac{\partial^4 f}{\partial y^4} = g(x,y) \quad (32)$$

is an elliptic equation of frequent occurrence in problems of elastic stress analysis. When $g = 0$ it is called the *biharmonic* equation. Associated with the differential equation we now need two conditions at each boundary point, the most common conditions specifying boundary values of both f and its normal derivative $\partial f/\partial \nu$. The approximate finite-difference form of $\nabla^4 f$ is given by

$$h^4\nabla^4 f_0 = 20f_0-8\sum_1^4 f_r+2\sum_5^8 f_r+\sum_9^{12} f_r, \quad (33)$$

in the notation of Figure 3, and the two boundary conditions again permit the production of a set of algebraic equations, equal in number to the number of internal mesh points.

The general equation contains thirteen unknown pivotal values. Iterative solution of the equations is more difficult than in the second-order case, mainly due to ill-conditioning, but direct methods [**110**], [**111**] based on matrix operations are quite practicable.

The problem of curved boundaries is in a sense less difficult in this case than for second-order equations. This is associated with the fact that a knowledge of both f and $\partial f/\partial \nu$ at the boundary implies a knowledge of both $\partial f/\partial x$ and $\partial f/\partial y$, the derivatives in the directions of the mesh lines.

20. Problems in elasticity may also involve the solution of two simultaneous second-order equations of the form

$$\left.\begin{aligned} A\frac{\partial^2 u}{\partial x^2}+B\frac{\partial^2 u}{\partial y^2}+C\frac{\partial^2 v}{\partial x\,\partial y} &= 0,\\ B\frac{\partial^2 v}{\partial x^2}+A\frac{\partial^2 v}{\partial y^2}+C\frac{\partial^2 u}{\partial x\,\partial y} &= 0,\end{aligned}\right\} \tag{34}$$

with u and v, or some equivalent conditions, imposed at all points of a closed boundary. By replacing derivatives by finite differences, in the usual way, we produce a set of algebraic equations in the unknown pivotal values of u and v. The only new feature is contained in the approximate equation

$$4h^2\frac{\partial^2 u_0}{\partial x\,\partial y} = u_5-u_6+u_7-u_8, \tag{35}$$

in the notation of Figure 3.

In all these problems the difference corrections have a known form, and can be introduced at a later stage to correct a first approximation.

21. Some elliptic equations, such as

$$\nabla^2 f+\lambda f = 0, \tag{36}$$

are of eigenvalue type, and lead to the solution of the algebraic problem

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{f} = 0, \tag{37}$$

already discussed in Chapter 3.

22. Finally, in some problems, notably in fluid motion and in plasticity, the position of part of the boundary is not known in advance, and an extra condition is imposed at this boundary to fix its position. The usual method is one of iteration: the problem is solved for guessed boundary positions, and the true position estimated to fit the extra condition.

13

EVALUATION OF LIMITS; USE OF RECURRENCE RELATIONS

1. In this chapter we discuss the evaluation of limits of sequences, slowly convergent series and continued fractions, and the use of recurrence relations. The emphasis is on numerical procedures of wide applicability rather than mathematical transformations relevant only to specialized problems.

RICHARDSON'S DEFERRED APPROACH TO THE LIMIT

2. This method can often be used to improve approximate results obtained by finite-difference methods, without the explicit use of a difference correction. The approximation, f_1 say, obtained with the neglect of the difference correction differs from the true value of f by an amount depending on the interval h_1. Let us suppose that for small h_1 the approximation f_1 is of the form

$$f_1 = f + Ah_1^k + Bh_1^{k+1} + \ldots, \tag{1}$$

where A and B are unknown constants. If we calculate two approximations, f_1 and f_2, using different intervals h_1 and h_2, we have two equations of the form (1), from which we can eliminate A. In this way we obtain a new approximation

$$\frac{h_2^k f_1 - h_1^k f_2}{h_2^k - h_1^k} = f + B\frac{h_2^k h_1^{k+1} - h_1^k h_2^{k+1}}{h_2^k - h_1^k} + \ldots, \tag{2}$$

which may be termed the h^k-*extrapolation* formula. In particular, if $k = 2$ and $h_2 = \frac{1}{2}h_1$, equation (2) becomes

$$f \fallingdotseq f_2 + \tfrac{1}{3}(f_2 - f_1), \tag{3}$$

with an error $\frac{1}{6}Bh_1^3 + \ldots$. The method can clearly be extended to obtain formulae which take account of further terms in the expansion (1). An investigation of the conditions under which the particular form (3) is valid in the solution of differential equations by finite-difference methods is reported in [**119**].

Generally speaking, h^k-extrapolation is applicable whenever a kth-order process is used to compute, for example, approximations to an integral or the solution of a differential equation, provided that there are no singularities in the range of integration. In practice it is usually desirable to employ at least three different intervals in order to allow comparison between the results of two or more applications of (2); this will help to ensure that the f_r have been computed to sufficient accuracy, that the intervals used are sufficiently small and that the correct value of k has been chosen. If the exponent is not known, it can be determined numerically (provided that an expansion of the form (1) is known to exist) from the approximate formula

$$2^k \doteq \frac{f_2-f_1}{f_3-f_2},$$

where the intervals corresponding to f_1, f_2, f_3 satisfy the relations $h_3 = \frac{1}{2}h_2 = \frac{1}{4}h_1$.

As an example, let us evaluate the integral $\int_0^{\frac{1}{2}\pi} \sin x\, dx = 1$ by applying h^2-extrapolation to approximations obtained with the use of the trapezoidal rule (Chapter 7, § 11). Using intervals of $\frac{1}{4}\pi, \frac{1}{6}\pi$ and $\frac{1}{8}\pi$, we obtain 0·94806, 0·97705, 0·98712 respectively; applying formula (2) to the first and last and also to the last two of these approximations, we derive the values 1·00014 and 1·00007 respectively, from which the correct value can be inferred to within a unit of the fourth decimal. The excellence of this result is due to the fact that the coefficient of h^3 in (1) (with $k = 2$) is in this case zero, while that of h^4 is small compared with A; these favourable circumstances are quite common in practice.

EXPONENTIAL EXTRAPOLATION

3. This device, which is also known as *Aitken's δ^2-process* [**39**], can often be used to accelerate the convergence of an infinite sequence or iterative process. If x_{r-1}, x_r, x_{r+1} are three successive approximations to a quantity x, and if the errors $x-x_r$ are approximately in geometric progression, a better approximation is provided by the expression

$$x_r' = \frac{x_{r-1}x_{r+1}-x_r^2}{x_{r+1}-2x_r+x_{r-1}} = x_{r+1} - \frac{(x_{r+1}-x_r)^2}{x_{r+1}-2x_r+x_{r-1}}. \qquad (4)$$

The two forms are equivalent; the second is often more convenient computationally as the limit is approached (see [**2**], § 3.4). The formula may also be used with x_{r-1}, x_{r+1} replaced by x_{r-p}, x_{r+p}, where p is an integer greater than unity.

The process can be extended by forming the sequence $x_r', x_{r+1}', \ldots$, applying (4) again to produce a further sequence $x_{r+1}'', x_{r+2}'', \ldots$, and so on.

In the special case when the form of the relation between each iterate x_r and its predecessor is independent of r, it is usually better to proceed as follows. For a given x_0 we perform two iterations to produce x_1 and x_2, and calculate x_1' according to (4). The cycle is then repeated with x_0 replaced by x_1', and so on. Because of the symmetry of formula (4), a meaningful result corresponding to $r = -\infty$ may sometimes be obtained in this way even if the given sequence diverges as $r \to +\infty$.

Example

4. The smallest positive root of the equation $\tan x = e^x$ can be computed iteratively using $x_{r+1} = \tan^{-1}(e^{x_r})$, with $x_0 = 0$, but the convergence is rather slow. The first three iterates derived in this way are

$$0{\cdot}78540, \qquad 1{\cdot}14302, \qquad 1{\cdot}26213.$$

Applying formula (4) to this triad and performing two more iterations we obtain

$$1{\cdot}32161, \qquad 1{\cdot}31016, \qquad 1{\cdot}30729,$$

and a further application of (4) yields the value 1·30633, which is already correct to five decimal places.

SUMMATION OF SLOWLY CONVERGENT SERIES

5. This problem is closely related to that of evaluating the limit of a sequence; indeed the two are mathematically equivalent. Many special transformations have been devised to deal with particular types of series, but most of the computer's practical needs are likely to be met either by applying the method of § 3 to the sequence of partial sums or by using one or other of the transformations of Euler and van Wijngaarden described in §§ 6 to 11 below.

It must be assumed that the terms in the infinite series ultimately conform to a regular pattern, for otherwise the value of the sum of the infinite series could not be inferred by considering only a finite number of terms. If, as frequently happens, the early terms are irregular or decrease fairly rapidly, they can be summed directly and the selected method of extrapolation applied to the remaining series. We may also observe in passing that a given series can sometimes be brought to a more tractable form by a simple rearrangement or regrouping of its terms.

THE EULER TRANSFORMATION

6. This transformation, given by

$$\sum_{s=0}^{\infty}(-)^s u_s = \sum_{s=0}^{\infty}\frac{(-)^s}{2^{s+1}}\Delta^s u_0, \tag{5}$$

is chiefly used to sum series whose terms alternate in sign. It is valid [**51**] whenever both series converge, and may be demonstrated by symbolic methods as follows:

$$\sum_{s=0}^{\infty}(-)^s u_s = \sum_{s=0}^{\infty}(-E)^s u_0 = (1+E)^{-1}u_0 = \tfrac{1}{2}(1+\tfrac{1}{2}\Delta)^{-1}u_0 = \tfrac{1}{2}\sum_{s=0}^{\infty}(-\tfrac{1}{2}\Delta)^s u_0. \tag{6}$$

7. As an example consider the series

$$1-\tfrac{1}{3}+\tfrac{1}{5}-\tfrac{1}{7}+\ldots = \tfrac{1}{4}\pi = 0{\cdot}78540\ldots. \tag{7}$$

Working to five decimal places, we find by straightforward addition that the sum of the first six terms is 0·74401. Let the remaining series be denoted by $\sum_{s=0}^{\infty}(-)^s u_s$. Then the terms in the transformed series (5) may be evaluated as follows:

s	u_s	Δ	Δ^2	Δ^3	Δ^4	$(-)^s\Delta^s u_0/2^{s+1}$
0	+0·07692					+0·03846
		−1025				
1	·06667		+240			·00256
		785		−74		
2	·05882		166		+26	·00030
		619		−48		
3	·05263		+118			·00005
		−501				
4	+ ·04762					·00001
						$\Sigma =$ +0·04138

For the complete sum we have

$$0{\cdot}74401+0{\cdot}04138 = 0{\cdot}78539.$$

8. The application of Euler's transformation is not restricted to convergent series; meaningful results can sometimes be obtained even when the original series diverges, and the method is widely used in the summation of asymptotic series. In such cases, however, the reliability of the results should, if possible, be tested by applying an independent numerical check.

9. A more general form of the transformation, given by

$$\sum_{s=0}^{\infty}(-x)^s u_s = \sum_{s=0}^{\infty}\frac{(-x)^s}{(1+x)^{s+1}}\Delta^s u_0, \tag{8}$$

is useful in summing power series for various values of x. Further, by substituting $x = e^{i\theta}$ we can derive formulae for transforming sine and cosine series.

10. A modification of the procedure of §§ 6 and 7, due to van Wijngaarden, is ideal for automatic computation. We express Euler's transformation in the form

$$S \equiv \sum_{s=0}^{\infty} v_s = v_0+v_1+\ \dots\ +v_{n-1}+\tfrac{1}{2}\sum_{s=0}^{\infty} M^s v_n, \tag{9}$$

where $M = \frac{1}{2}(1+E)$ is the *forward mean operator* defined by

$$Mv_s = \tfrac{1}{2}(v_s+v_{s+1}). \tag{10}$$

The equivalence of (5) and (9), with $v_{n+s} = (-)^s u_s$, may easily be verified. We denote by $S_{n,p}$ the approximation to S obtained when the upper limit on the right of (9) is replaced by p.

Starting with $S_{0,0} = \frac{1}{2}v_0$, we take as our next approximation $S_{0,1} = \frac{1}{2}(v_0+Mv_0)$ or $S_{1,0} = v_0+\frac{1}{2}v_1$, according as $\frac{1}{2}Mv_0$ or $\frac{1}{2}v_1$ is smaller; to obtain $S_{0,1}$ or $S_{1,0}$ we add either $\frac{1}{2}Mv_0$ or Mv_0 to $S_{0,0}$. In general, either p or n is increased by unity at each step, according as $M^{p+1}v_n$ or $M^p v_{n+1}$

is the smaller, and a new partial sum $S_{n,p+1}$ or $S_{n+1,p}$ obtained by adding $\frac{1}{2}M^{p+1}v_n$ or $M^{p+1}v_n$ to $S_{n,p}$. A useful practical criterion for terminating the calculation is the negligibility of the added term in two (or more) consecutive cycles.

An important feature of the method is the economy of storage which it permits. At any stage it is necessary to retain only a single row of backward means, typified by

$$v_{n+p}, \quad Mv_{n+p-1}, \quad M^2 v_{n+p-2}, \quad \ldots, \quad M^p v_n. \tag{11}$$

During the cycle which follows the calculation of $S_{n,p}$, this row is replaced by

$$v_{n+p+1}, \quad Mv_{n+p}, \quad M^2 v_{n+p-1}, \quad \ldots, \quad M^{p+1} v_n; \tag{12}$$

if n, and not p, is increased, the last term is subsequently dropped. As soon as the kth member of (11) has contributed to forming the $(k+1)$th member of (12), it can be overwritten by the kth member of (12).

VAN WIJNGAARDEN'S TRANSFORMATION

11. The powerful Euler technique is not directly applicable to series of positive terms. However, by means of a transformation due to van Wijngaarden, given by

$$S \equiv \sum_{r=1}^{\infty} v_r = \sum_{r=1}^{\infty} (-)^{r-1} w_r, \tag{13}$$

where

$$w_r = v_r + 2v_{2r} + 4v_{4r} + 8v_{8r} + \ldots, \tag{14}$$

we can convert a series of positive terms into an alternating series, to which Euler's transformation can then be applied.

From (14) we deduce the relation

$$2w_{2s} = w_s - v_s, \tag{15}$$

which can be used to compute the w_s of even suffix. It also provides the basis for a simple proof of (13); thus

$$\begin{aligned} v_1 + v_2 + v_3 + \ldots &= (w_1 - 2w_2) + (w_2 - 2w_4) + (w_3 - 2w_6) + \ldots \\ &= w_1 - w_2 + w_3 - w_4 + \ldots. \end{aligned} \tag{16}$$

The conditions needed to justify the rearrangement and ensure the convergence of (14) are very mild; for example, it is sufficient that either $|v_{r+1}| \leqslant |v_r|$ and Σv_r converges, or that $|v_r| \leqslant Kr^{-c-1}$, where K and c are positive constants.

As an example, if $v_r = r^{-c-1}$ $(c > 0)$, we have

$$\begin{aligned} w_r &= r^{-c-1} + 2(2r)^{-c-1} + 4(4r)^{-c-1} + \ldots \\ &= r^{-c-1}(1 + 2^{-c} + 4^{-c} + \ldots) = r^{-c-1}/(1 - 2^{-c}). \end{aligned}$$

Hence
$$S = \frac{1}{1-2^{-c}}\left(\frac{1}{1^{1+c}} - \frac{1}{2^{1+c}} + \frac{1}{3^{1+c}} - \ldots\right).$$

COMPARISON WITH AN INTEGRAL

12. If v_r is the value at $x = r$ of a function $v(x)$ whose integral is known, an integration formula can be used to sum the series. For example, we may employ the central-difference quadrature formula (25) of Chapter 7 in the form

$$\tfrac{1}{2}v_n + \sum_{r=n+1}^{\infty} v_r = \int_n^{\infty} v(x)\,dx - (\tfrac{1}{12}\mu\delta - \tfrac{11}{720}\mu\delta^3 + \dots)\,v_n, \tag{17}$$

or the *Euler–Maclaurin formula*

$$\tfrac{1}{2}v_n + \sum_{r=n+1}^{\infty} v_r = \int_n^{\infty} v(x)\,dx - \{\tfrac{1}{12}v'(n) - \tfrac{1}{720}v'''(n) + \dots\}. \tag{18}$$

Even if the given series cannot immediately be dealt with in this way, it may yet be possible to subtract from each term v_r a quantity v_r^* which approaches v_r asymptotically for large r and is such that $\int_n^{\infty} v^*(x)\,dx$ can be easily evaluated or found from tables. The residual series $\Sigma(v_r - v_r^*)$ then converges more rapidly than the original series and it may be practicable to sum it directly.

EVALUATION OF CONTINUED FRACTIONS

13. Another type of limiting expression of fairly common occurrence is the infinite continued fraction, obtained by letting n tend to infinity in the expression

$$b_0 + \frac{a_1}{b_1 +}\,\frac{a_2}{b_2 +}\,\cdots\,\frac{a_n}{b_n}. \tag{19}$$

(For a discussion of convergence and an account of the general theory, see [**125**] or [**126**].)

The expression (19) is known as the *nth approximant* or *convergent*. It can be calculated directly by alternate division and addition, working backwards from the right; in this case a check must be applied to ensure that the value of n chosen is sufficiently large. Alternatively, the successive approximants A_n/B_n may be generated by means of the recurrence relations

$$A_{n+1} = b_{n+1}A_n + a_{n+1}A_{n-1}, \qquad B_{n+1} = b_{n+1}B_n + a_{n+1}B_{n-1}, \tag{20}$$

with $A_0 = b_0$, $\quad A_1 = b_0 b_1 + a_1$, $\quad B_0 = 1$, $\quad B_1 = b_1$,

or by use of the summation formula

$$\frac{A_n}{B_n} = b_0 + \sum_{i=1}^{n} \rho_1 \rho_2 \cdots \rho_i, \tag{21}$$

where the ρ_i are given by

$$r_i = \frac{a_i}{b_{i-1}b_i}, \quad \rho_1 = \frac{a_1}{b_1}, \quad 1 + \rho_2 = \frac{1}{1 + r_2}, \quad 1 + \rho_i = \frac{1}{1 + r_i(1 + \rho_{i-1})} \quad (i \geqslant 3). \tag{22}$$

In automatic work, the second method has the slight disadvantage that the numbers A_n and B_n are apt to grow large. By contrast, the number of figures in the ρ_i can actually be reduced as the terms in the sum (21) decrease; but the recurrence (22) requires modification if any of the b_i vanish or are very small.

THE USE OF RECURRENCE RELATIONS

14. Recurrence relations are often of assistance in the computation of mathematical functions; for example, the recurrence relations satisfied by Bessel functions were widely used in the tabulation of these functions. They give rise to a very economical form of computing, since each step in the computation yields one required value of the function.

Careful attention should be given to the possibility of error build-up when recurrence relations are used. The dependence of the error in each newly computed value on those of its immediate predecessors is usually apparent by inspection, but since these errors are interrelated the overall pattern is not immediately obvious. In the case of a *linear* recurrence relation, the most common in practice, the errors themselves represent a solution; the way in which they are propagated can then be easily determined by applying the recurrence with arbitrary starting values and observing the rate at which the resulting solution increases.

15. Consider the relation

$$f_{n+1}-\frac{2n}{x}f_n+f_{n-1}=0, \tag{23}$$

satisfied by the Bessel functions J_n and Y_n. If J_0 and J_1 are known, (23) can be used to calculate $J_2, J_3, \ldots$ successively, all for the same argument x. The process is accurate, however, only so long as n does not exceed x; thereafter there is a rapid build-up of error. This is because the solution obtained inevitably contains a small multiple of the unwanted solution Y_n, which increases exponentially with n when $n>x$, whereas the wanted function J_n decreases. On the other hand, if we calculate Y_n by the same procedure, with Y_0 and Y_1 given, the wanted function increases as fast as the error and there is no loss of significant figures.

16. An alternative method for calculating J_n when $n>x$ is to choose N so large that J_N is negligible and recur *backwards*, taking as initial conditions $f_{N+1}=0, f_N=1$. In this case Y_n decreases while J_n increases, and the unwanted part of the solution decays. Thus the values obtained are effectively those of J_n multiplied by a constant factor; this can finally be determined from some known value of J, for example $J_0(x)$, or independently by using the relation

$$J_0+2J_2+2J_4+\ldots=1.$$

The same principle is employed in the method for calculating Chebyshev coefficients described in Chapter 9, § 23.

17. Useful information can sometimes be obtained by considering the limiting form of the recurrence relation for large n. For example, when $n\to\infty$ the equation

$$(n+1)f_{n+1}-(2n+1)xf_n+nf_{n-1}=0, \tag{24}$$

satisfied by the Legendre functions $P_n(x)$ and $Q_n(x)$, assumes the limiting form

$$f_{n+1} - 2xf_n + f_{n-1} = 0. \tag{25}$$

This is a difference equation with constant coefficients with the general solution

$$\text{(i)} \qquad Ae^{in\theta} + Be^{-in\theta} \quad \text{with} \quad \theta = \cos^{-1} x, \tag{26}$$

$$\text{or} \quad \text{(ii)} \qquad Ae^{n\alpha} + Be^{-n\alpha} \quad \text{with} \quad \alpha = \cosh^{-1} x, \tag{27}$$

according as x is less than or greater than unity; for simplicity we consider only positive values of x.

In case (i) all solutions of (25) are bounded as $n \to \infty$. This suggests that in the evaluation of any particular solution of (24) by recurrence there is scarcely any tendency for the error to grow. (We disregard the mere accumulation of rounding errors, which is usually of secondary significance.) In case (ii), however, the solution (27) indicates that equation (24) possesses solutions of exponential type for large n; by analogy with the computation of J_n and Y_n discussed in §§ 15 and 16, the direction in which the recurrence should proceed will thus depend on whether the required solution increases or decreases with n.

14

EVALUATION OF INTEGRALS

GENERAL METHODS

1. The numerical evaluation of an integral on *desk machines* is usually carried out using one of the finite-difference quadrature formulae of Chapter 7. These formulae are particularly well suited to the evaluation of an indefinite integral at several successive tabular points. Indeed their use is often preferred even when a closed expression is available for the integral.

2. As an example, consider the integral

$$\int_0^x \frac{t}{a^3+t^3}\,dt = \frac{1}{6a}\log\frac{x^2-ax+a^2}{(x+a)^2}+\frac{1}{a\sqrt{3}}\tan^{-1}\frac{x\sqrt{3}}{2a-x} \quad (-a<x<2a). \quad (1)$$

If this is required for a single value of x, computation of the analytical expression is a suitable method. For a sequence of values of x, however, it is much easier to compute the integral by quadrature using, for example, formula (27) of Chapter 7. The analytical formula will nevertheless still be used to provide spot checks for one or more values of x.

3. In *automatic work*, the more complicated finite-difference formulae are usually discarded in favour of *Simpson's rule* (Chapter 7, equation (34)).

In the case of definite integration, the result of repeatedly applying this formula may be expressed in the form

$$\int_0^1 y\,dx = \tfrac{1}{3}h\{y(0)+y(1)+2S_h+4T_h\}, \quad (2)$$

where

$$S_h = y(2h)+y(4h)+\ldots+y\{(n-2)h\},$$
$$T_h = y(h)+y(3h)+\ldots+y\{(n-1)h\},$$

and $n = 1/h$ is an even positive integer. The interval h must be sufficiently small to ensure that the truncation error is negligible. A convenient automatic procedure is to apply the formula with a coarse interval initially, then to halve it repeatedly until the results of two (or more) successive applications agree. Only alternate ordinates have to be computed and summed each time the interval is halved; $S_{\frac{1}{2}h}$ is obtained by adding S_h and T_h, both of which are available from the previous stage.

For indefinite integration, unless an interval size which is both safe and economical can be determined beforehand it is desirable to test at every step whether the current interval should be halved or may be doubled.

For further details of the methods of this and the following two sections, see [**134**].

4. Another suitable method, described in Chapter 8, § 19, is the termwise integration of the Chebyshev expansion of the integrand; this is particularly valuable for indefinite integration.

5. A third general method which is well suited to the automatic evaluation of definite integrals is the use of the *Gauss quadrature formula.* The transformation

$$x = \tfrac{1}{2}(a+b)+\tfrac{1}{2}(b-a)X$$

yields the standard form

$$\int_a^b y\,dx = \tfrac{1}{2}(b-a)\int_{-1}^{1} y\,dX = \tfrac{1}{2}(b-a)\sum_{r=1}^{n} w_r^{(n)}\,y(X_r^{(n)}), \tag{3}$$

where the points $X_r^{(n)}$ and the weights $w_r^{(n)}$ are chosen so that the formula is exact when y is any polynomial of degree less than $2n$. Thus by the use of n points the Gauss formula achieves an accuracy comparable with that of an equal-interval formula using $2n$ points.

This advantage is to some extent offset by the increased difficulty of checking. In order to ensure that the error associated with a particular value of n is negligible, it may be necessary to repeat the calculation using a different n, and this entails the evaluation of a completely new set of ordinates. For this reason, the method achieves the greatest economy when applied to a batch of similar integrals for which a suitable value of n can be determined by examining one or two test cases.

Extensive tables of $X_r^{(n)}$ and $w_r^{(n)}$ are given in [**137**], [**138**] and [**139**].

INFINITE INTEGRALS

6. When the upper limit of integration is infinite, the finite-difference formula (25) of Chapter 7 yields

$$\int_{x_0}^{\infty} y\,dx = h(\tfrac{1}{2}y_0+\sum_{r=1}^{\infty} y_r+Cy_0), \tag{4}$$

where $y_r = y(x_0+rh)$ and

$$Cy_0 = (\tfrac{1}{12}\mu\delta-\tfrac{11}{720}\mu\delta^3+\ldots)\,y_0. \tag{5}$$

If the convergence of the integral is rapid, the expression (4) is readily evaluated and no new problem arises. As with finite integrals, it will usually be convenient to choose h so small that only a few terms in the difference correction (5) are needed. In the special case when the mean odd differences at x_0 (and hence Cy_0) vanish, however, it is practicable to use a comparatively large interval. This happens, for example, when $x_0 = -\infty$, in which case (4) reduces to the trapezoidal rule:

$$\int_{-\infty}^{\infty} y\,dx = h\sum_{r=-\infty}^{\infty} y(rh). \tag{6}$$

The accuracy of a result obtained from (6) can, as a rule, be checked by repetition using a different h, the use of differences of y being thus altogether avoided. The formula is not, of course, exact; its asymptotic nature is discussed and bounds for the error determined by Goodwin [**141**] in the case of integrands of the form $e^{-x^2}f(x)$. See also [**142**].

7. The *Laguerre–Gauss* and *Hermite–Gauss formulae,*

$$\int_0^\infty e^{-x}y(x)\,dx = \sum_{r=1}^{n} \alpha_r^{(n)}\,y(x_r^{(n)}), \qquad \int_{-\infty}^{\infty} e^{-x^2}y(x)\,dx = \sum_{r=1}^{n} \bar{\alpha}_r^{(n)}\,y(\bar{x}_r^{(n)}), \tag{7}$$

are sometimes useful when the arbitrary function $y(x)$ approximates to a polynomial of low or moderate degree. For tables of the requisite abscissae and weights see [**137**] and [**140**].

8. In considering the problem of the evaluation of *slowly convergent integrals,* two principal cases may be distinguished: (a) the integrand $y(x)$ decreases steadily to zero as $x \to \infty$, (b) $y(x)$ oscillates in a regular manner about zero as $x \to \infty$. Both cases may often be dealt with by first expressing the integral as an infinite series of ordinates together with a difference correction, as in equation (4), and then applying one of the methods for summing slowly convergent series described in Chapter 13.

9. In case (a), the methods of Aitken and van Wijngaarden (Chapter 13, §§ 3 and 11) are directly applicable to the series $\sum\limits_{r=1}^{\infty} y_r$.

In case (b), let us suppose that for sufficiently large r the sign of y_r changes approximately every m terms. We first group the terms in the form

$$\begin{aligned}\sum_{r=1}^{\infty} y_r &= (y_1 + \dots + y_m) + (y_{m+1} + \dots + y_{2m}) + (y_{2m+1} + \dots + y_{3m}) + \dots \\ &= z_1 + z_2 + z_3 + \dots, \quad \text{say,}\end{aligned} \tag{8}$$

and then apply Euler's transformation (Chapter 13, §§ 6 to 10). A numerical example is presented in § 12 below.

10. This method is simple and effective, but may require a large number of ordinates y_r. An alternative procedure for evaluating oscillatory integrals, which demands the computation of fewer ordinates, is as follows. First, consider integrals of the form $\int_{n\pi}^{\infty} f(x)\sin x\,dx$, where $f(x)$ is a steadily decreasing function whose differences at the interval π are well-behaved in the range of integration. Substituting for f in the sub-range $r\pi \leqslant x \leqslant (r+1)\pi$ in terms of Everett's formula (Chapter 7, § 7) and integrating term by term, we derive an expansion of the form

$$\int_{r\pi}^{(r+1)\pi} f(x)\sin x\,dx = (-)^r(1 + a_1\delta^2 + a_2\delta^4 + \dots)(f_r + f_{r+1}), \tag{9}$$

where $f_r = f(r\pi)$. Hence, summing from $r = n$ to $r = \infty$ we obtain

$$\int_{n\pi}^{\infty} f(x)\sin x\,dx = (-)^n(1 + a_1\delta^2 + a_2\delta^4 + \dots)f_n. \tag{10}$$

The first few coefficients a_i, obtained by expanding the operator $\{1 + (\log E)^2/\pi^2\}^{-1}$ in powers of δ^2 (Chapter 7, § 3), are

$$a_1 = -0{\cdot}10132\,118, \qquad a_2 = 0{\cdot}01870\,941,$$
$$a_3 = -0{\cdot}00387\,695, \qquad a_4 = 0{\cdot}00084\,579.$$

11. The method is readily extended to the case in which the integrand is expressible in the form $M \sin\theta$, where the *modulus* M varies slowly and the *phase* θ increases steadily with x. We then take θ as a new variable of integration and write

$$\int_{x_0}^{\infty} M \sin\theta \, dx = \int_{\theta_0}^{\infty} \frac{M \sin\theta}{d\theta/dx} \, d\theta, \tag{11}$$

which is amenable to the treatment described.

Example

12. Let us apply the methods of §§ 9 to 11 to evaluate

$$\int_{x_0}^{\infty} J_0(x) \, dx.$$

Taking $x_0 = 0$, $h = 1$ and $m = 3$, and using the first thirty ordinates, we obtain for the series (8)

$$\sum_{r=1}^{\infty} z_r = 0{\cdot}72904 - 0{\cdot}42410 + 0{\cdot}38140 - 0{\cdot}36944 + 0{\cdot}36378 - 0{\cdot}35811$$
$$+ 0{\cdot}35023 - 0{\cdot}33929 + 0{\cdot}32501 - 0{\cdot}30738 + \dots.$$

Summing the first three terms directly and applying Euler's transformation to the remainder, we obtain $\sum\limits_{r=1}^{\infty} z_r = 0{\cdot}50001$. Application of formula (4), with $y_0 = 1, Cy_0 = 0$, then yields for the integral the value 1·00001, in virtual agreement with the true value, unity.

The modulus-phase representation required for the application of the second method is here given by

$$J_0(x) = M \cos\psi, \qquad M^2 = J_0^2 + Y_0^2, \qquad M^2 \frac{d\psi}{dx} = \frac{2}{\pi x} \tag{12}$$

(compare Chapter 15, § 20). Hence taking $x_0 = j_{0,n}$, the nth zero of $J_0(x)$, we obtain

$$\int_{j_{0,n}}^{\infty} J_0(x) \, dx = \int_{(n-\frac{1}{2})\pi}^{\infty} \frac{M \cos\psi}{d\psi/dx} \, d\psi = \int_{n\pi}^{\infty} (\tfrac{1}{2}\pi x M^3) \sin\theta \, d\theta,$$

where $\theta = \psi + \frac{1}{2}\pi$. Hence from (10)

$$\int_{j_{0,n}}^{\infty} J_0(x) \, dx = (-)^n (1 + a_1 \delta^2 + a_2 \delta^4 + \dots)(\tfrac{1}{2}\pi x M^3)_n, \tag{13}$$

where the differences are those of $\frac{1}{2}\pi x M^3$ tabulated at an interval π in ψ, and the suffix n refers to the point $\psi = (n - \frac{1}{2})\pi$, corresponding to $x = j_{0,n}$. With $n = 5$, for example, application of (13) yields, with the use of seven ordinates,

$$\int_{j_{0,5}}^{\infty} J_0(x) \, dx = -0{\cdot}20565,$$

which is correct to five decimal places.

13. Many types of singularity can be removed by a suitable *change of variable*. For example, integrals of the forms

$$\int_0^1 f(x)\,x^{-\frac{1}{2}}\,dx, \qquad \int_0^1 f(x)\ln x\,dx, \qquad \int_{-1}^1 f(x)\,(1-x^2)^{-\frac{1}{2}}\,dx,$$

in which $f(x)$ is well behaved, become non-singular with the transformations

$$x = t^2, \qquad x = e^{-t}, \qquad x = \sin t,$$

respectively. In the second case, one limit of integration becomes infinite but the convergence of the transformed integral is rapid. For the third integral, an alternative is to split the range at $x = 0$ and apply the algebraic transformations $x = \pm(1-u^2)$ to the upper and lower parts respectively.

14. Another approach is to employ a formula of the type

$$\int_a^b f(x)\,w(x)\,dx = \Sigma c_r f(x_r), \tag{14}$$

in which $f(x)$ is again an arbitrary well-behaved function, the points x_r are prescribed, and the coefficients depend on the given function $w(x)$ (which may be singular in $a \leqslant x \leqslant b$) and on the limits a, b.

When the points x_r are equally spaced, the formula is said to be of *Newton–Cotes type* (compare Chapter 7, § 16); the case $w(x) = \ln x$ is treated in [**143**], and the three cases $w(x) = x^{\pm\frac{1}{2}}$ and $\ln x$ in [**144**]; see also [**188**].

If the abscissae are chosen to secure the maximum accuracy consistent with a given order of formula, we have a formula of *Gauss type* (compare §§ 5 and 7); examples are the *Chebyshev–Gauss* and *Jacobi–Gauss* quadrature formulae corresponding to the weight functions

$$w(x) = (1-x^2)^{-\frac{1}{2}} \quad \text{and} \quad w(x) = (1-x)^{\alpha}(1+x)^{\beta} \quad (\alpha > -1, \beta > -1)$$

respectively, and limits of integration $a = -1$, $b = 1$. For details see [**4**] and [**5**].

15. The method of the *extraction of a singular part* may be illustrated by considering the integral

$$I(x) = \int_0^x \frac{e^{-u}}{1-u}\,du. \tag{15}$$

We note that the integrand has a pole at $u = 1$ and that $I(1) = \infty$. In the neighbourhood of $u = 1$ the integrand behaves like $e^{-1}(1-u)^{-1}$; accordingly we express $I(x)$ in the form

$$I(x) = e^{-1}\int_0^x \frac{du}{1-u} + \int_0^x \frac{e^{-u}-e^{-1}}{1-u}\,du.$$

The first term on the right is equal to $-e^{-1}\ln|1-x|$, while the second has no singularity at $x = 1$ and can be evaluated quite easily by numerical quadrature. If $x > 1$ the value obtained in this way is the *Cauchy principal value* of $I(x)$.

16. While the discussion of methods of a purely mathematical character is beyond the scope of this manual, mention may be made of a few basic mathematical procedures which frequently facilitate the numerical evaluation of integrals. Further details are given in [**135**].

17. A fairly common requirement is the integration for $n = 0, 1, 2, \ldots$ of a function containing a factor such as $x^n, \cos nx$ or $J_n(x)$. Integrals of these types may satisfy a linear *recurrence relation.*

A simple example is afforded by the generalized exponential integral $E_n(x)$. Integrating by parts we obtain

$$E_n(x) = \int_1^\infty \frac{e^{-xu}}{u^n} du = \left[\frac{-e^{-xu}}{(n-1)u^{n-1}}\right]_1^\infty - \frac{x}{n-1}\int_1^\infty \frac{e^{-xu}}{u^{n-1}} du$$

$$= \frac{1}{n-1}\{e^{-x} - xE_{n-1}(x)\}. \qquad (16)$$

From the values of the exponential integral $E_1(x) = -\mathrm{Ei}(-x)$, we can compute $E_2(x), E_3(x), \ldots$ in succession by means of (16); many guarding figures are needed, however, when x is large (see Chapter 13, § 14).

18. If a function defined by a definite integral satisfies a *differential equation* with respect to a parameter, the most convenient way of evaluating the integral for a range of parameter values is often provided by the numerical integration of the differential equation. A value of the integral is thereby obtained at each step of the process of solution.

For example, the function

$$F(x) = \int_0^\infty \frac{e^{-u^2}}{u+x} du \qquad (17)$$

satisfies the differential equation

$$F'(x) + 2xF(x) = \sqrt{\pi} - (1/x). \qquad (18)$$

This equation was used by Goodwin and Staton [**136**] to compute $F(x)$ on desk machines. For automatic work, (18) could be used to derive Chebyshev expansions of $F(x)$ by the method of Chapter 9, § 23.

19. Although an integral may be amenable to expansion in a variety of ways, it is generally advisable to consider a purely numerical approach before undertaking a lengthy mathematical investigation. Nevertheless, a *series expansion* will often afford a convenient means of computation, particularly in the neighbourhood of a singularity. For example, the integral (17) can be computed for small values of x by means of the ascending series

$$F(x) = -e^{-x^2}\ln x + e^{-x^2}\left[\sqrt{\pi}\sum_{n=0}^\infty \frac{x^{2n+1}}{n!(2n+1)} - \sum_{n=1}^\infty \frac{x^{2n}}{n!\,2n} - \frac{\gamma}{2}\right], \qquad (19)$$

where $\gamma = 0{\cdot}577\ldots$ is Euler's constant.

20. *Asymptotic expansions* are frequently used in computing integrals. If the given integral can be reduced to the form

$$\int_0^\infty e^{-xt}\phi(t)\,dt, \qquad (20)$$

then an expansion in descending powers of x may be obtained by expanding $\phi(t)$ in ascending powers of t and integrating formally term by term. Precise conditions justifying this procedure are given, for example, in [**145**]. Thus for the integral (17) we have

$$\begin{aligned} F(x) &= \frac{1}{2}\int_0^\infty \frac{e^{-t}}{t^{\frac{1}{2}}+x}\frac{dt}{t^{\frac{1}{2}}} \\ &= \frac{1}{2}\int_0^\infty \sum_{r=0}^\infty \frac{(-)^r t^{\frac{1}{2}r-\frac{1}{2}}}{x^{r+1}} e^{-t}\,dt \sim \frac{1}{2}\sum_{r=0}^\infty \frac{(-)^r \Gamma(\frac{1}{2}r+\frac{1}{2})}{x^{r+1}}. \end{aligned} \tag{21}$$

The useful computational range of this expansion may be increased by application of the Euler transformation (Chapter 13, § 8); see [**136**].

15

TABULATION OF MATHEMATICAL FUNCTIONS

INTRODUCTION

1. The preparation of a new numerical table of a mathematical function takes place in several stages. For convenience we may separate these into two main groups.

First, there is the mathematical and numerical investigation of the properties of the function to determine the most convenient method or methods of computation, and the computation itself.

Second, there is the checking, subtabulation (if required), preparation of interpolation aids, preparation of final copy and printing. These matters, which may be termed *principles of table-making*, form the main subject of this chapter.

Before discussing these topics, however, we consider a most important question confronting the table-maker: has the right choice of tabular functions been made? A type of difficulty encountered in making the correct choice is described in the next two sections.

CHOICE OF SOLUTIONS OF DIFFERENTIAL EQUATIONS

2. Consider, for example, the equation

$$\frac{d^2y}{dx^2} = y. \tag{1}$$

This has the general solution

$$y = Ae^x + Be^{-x}, \tag{2}$$

where A, B are arbitrary constants. Another form of the general solution is

$$y = A\cosh x + B\sinh x. \tag{3}$$

Given tables of e^x and e^{-x} having a prescribed number of figures, we can evaluate the expression (2) to a certain accuracy for any values of A, B and x. This accuracy is not attainable for large values of x if we use instead tables of $\cosh x$ and $\sinh x$ and the expression (3), because the leading figures of the corresponding values of these functions are the same, and this results in severe cancellation when A is approximately equal to $-B$.

For this reason, e^x and e^{-x} are said to be a *numerically satisfactory* pair of solutions of equation (1) for large values of x. The pair $\cosh x$ and $\sinh x$ are not numerically satisfactory, even though they are linearly independent in the mathematical sense.

This kind of difficulty is associated with differential equations whose solutions are of exponential type for large values of the independent variable, or are unbounded in the neighbourhood of a singularity. The choice between oscillatory solutions is usually far less critical.

3. A less simple example is provided by Bessel's equation

$$\frac{d^2y}{dx^2}+\frac{1}{x}\frac{dy}{dx}+y = 0. \tag{4}$$

In the neighbourhood of the singularity $x = 0$, the solution $J_0(x)$ is bounded and the solution $Y_0(x)$ unbounded; both functions oscillate for large positive x. Accordingly, $J_0(x)$ and $Y_0(x)$ comprise a numerically satisfactory pair of solutions for all real positive values of x.

In the complex plane, however, this is no longer true; both $J_0(z)$ and $Y_0(z)$ become exponentially large as z tends to infinity along any ray not parallel to the real axis. A numerically satisfactory pair of solutions in the upper half of the complex plane, both in the neighbourhood of the origin and at infinity, is $J_0(z)$ and the Hankel function $H_0^{(1)}(z)$.

PREPARATION OF PRINTED TABLES

4. When the function values have been computed they must be checked systematically. This is usually performed by differencing, using either an accounting machine with several registers or, if the values are already punched on cards, a punched-card tabulating machine. Any blunders that have been made in the computation will then be revealed. Small end-figure errors of a unit or so will not always be found in this way, so that the original computations and the differencing are performed with the retention of one or more guarding decimals, and subsequently rounded mechanically to the number of decimals required in the final table.

5. At this stage any *subtabulation* which is necessary will be carried out. If an automatic computer has been used for the calculations it is probable that every value required in the final table will have been evaluated directly. On desk machines, however, it is usually economic to perform the original calculations at the largest convenient interval in the argument at which the function has convergent differences. The intermediate values can then be filled in by systematic interpolation, the whole process being carried out mechanically by punched-card or accounting machine.

6. The final stages in the preparation of the table depend on the method of reproduction. In the past, tables were invariably set in type by compositors and printed in the usual way by letterpress. This necessitates careful and laborious checking of proofs to ensure that the correct figures have been printed. The proof-reading is carried out by comparison with the computed values, and by differencing, the latter method being much the sounder.

At present, increasing numbers of tables are reproduced photographically. This has the advantages of being slightly cheaper and reducing the amount of necessary proof-reading. Good copy can be prepared directly from the computer output or from punched cards by means of an automatic typewriter or line printer. Letterpress, however, has greater flexibility in arrangement and a more pleasing appearance, and many fundamental tables are still printed in this way.

INTERPOLATION AIDS

7. Many tables of elementary functions, such as logarithms and sines and cosines, are produced with an interval in the argument sufficiently small to ensure that linear interpolation is accurate. The compiler need then make no special provision for interpolation, except possibly to provide a table of mean first differences or proportional parts.

In the case of tables of higher mathematical functions, with more than three or four figures, economic and other limitations on space may be too stringent to permit provision of a linearly interpolable table. In this case the user must perform the interpolation by means of a more complicated formula, and it is incumbent on the compiler to ease the labour of interpolation by providing quantities in the table in addition to the function values. These quantities are known as *interpolation aids*.

Differences

8. The commonest aids are the central differences of the function, for use either with Bessel's formula

$$f_p = f_0 + p\delta_{\frac{1}{2}} + B_2(\delta_0^2 + \delta_1^2) + B_3\delta_{\frac{1}{2}}^3 + B_4(\delta_0^4 + \delta_1^4) + \dots, \tag{5}$$

or with Everett's formula

$$f_p = (1-p)f_0 + pf_1 + E_2\delta_0^2 + F_2\delta_1^2 + E_4\delta_0^4 + F_4\delta_1^4 + \dots; \tag{6}$$

see Chapter 7, §§ 7, 8. Everett's formula has the advantage that it does not use differences of odd order, and in consequence only those of even order need be given in the table.

The routine application of (5) or (6) requires tables of the interpolation coefficients $B_2, B_3, E_2, F_2, \dots$. *Interpolation and allied tables* [**167**] gives these and other coefficients and a detailed explanation of their use.

Modified differences

9. The power of differences is greatly increased by use of a device known as the *throw-back*, the basis of which is as follows. An examination of numerical tables of the interpolation coefficients $B_4(p)$ and $B_2(p)$ reveals that their ratio varies over the comparatively small range $(-\frac{1}{6}, -\frac{3}{16})$ when $0 \leqslant p \leqslant 1$. This suggests that we may allow for most of the effect of fourth differences by forming a *modified* second difference

$$\delta_m^2 = \delta^2 - C\delta^4, \tag{7}$$

where C is a constant. Neglecting differences of the fifth and higher orders, we can then write (5) in the form

$$f_p = f_0 + p\delta_{\frac{1}{2}} + B_2(\delta_{m0}^2 + \delta_{m1}^2) + B_3\delta_{\frac{1}{2}}^3, \tag{8}$$

with a residual error of amount

$$\epsilon = (B_4 + CB_2)(\delta_0^4 + \delta_1^4). \tag{9}$$

The constant C is chosen so that the coefficient in (9) has the smallest maximum numerical value in the range $0 \leqslant p \leqslant 1$. We find the approximate value 0·184 for C, and the maximum value of $|B_4 + 0{\cdot}184B_2|$ does not exceed 0·00023. Therefore if fourth differences do not exceed 1100 the error of this approximation is less than half a unit in the last decimal.

Because of the equivalence of Bessel's and Everett's formulae we can replace (6), in the same circumstances, by

$$f_p = (1-p)f_0 + pf_1 + E_2\delta_{m0}^2 + F_2\delta_{m1}^2. \tag{10}$$

Thus the effect of fourth differences which are less than 1100 can be allowed for by giving δ_m^2 in place of δ^2 in the table. The modified difference is treated in the same way as an ordinary difference in carrying out the interpolation. This device helps the compiler by removing the need to tabulate fourth differences, and the user by providing a simpler formula.

10. The idea can be extended. The formula

$$f_p = (1-p)f_0 + pf_1 + E_2\delta_{m0}^2 + F_2\delta_{m1}^2 + M_4\gamma_0^4 + N_4\gamma_1^4, \tag{11}$$

in which

$$\left.\begin{aligned} \delta_m^2 &= \delta^2 - 0{\cdot}184\delta^4 + 0{\cdot}03808\,2\delta^6 - 0{\cdot}00830\delta^8 + 0{\cdot}0019\delta^{10} - \ldots, \\ \gamma^4 &= 0{\cdot}001\delta^4 - 0{\cdot}00027\,83\delta^6 + 0{\cdot}00006\,8\delta^8 - 0{\cdot}00002\delta^{10} + \ldots, \end{aligned}\right\} \tag{12}$$

and

$$M_4 = 1000(E_4 + 0{\cdot}184E_2), \qquad N_4 = 1000(F_4 + 0{\cdot}184F_2), \tag{13}$$

allows for the effect of all differences, provided that

$$|\mu\delta^6| < 300{,}000 \quad \text{and} \quad |\delta^7| < 27{,}000.$$

These conditions ensure that the truncation error does not exceed one half-unit of the last figure given provided, of course, that the differences do not diverge (Chapter 7, § 17).

The full theory of 'throw-back' interpolation is given in [**150**].

Reduced derivatives

11. If we define

$$\tau^s \equiv h^s f^{(s)}(a)/s! \quad (s = 1, 2, \ldots), \tag{14}$$

where f is the tabulated function and h the argument interval, then the Taylor series at the tabular point $x = a$ may be written in the form

$$f(a+ph) = f(a) + p\tau + p^2\tau^2 + p^3\tau^3 + \ldots. \tag{15}$$

In this method the aids to interpolation are the quantities $\tau, \tau^2, \ldots$, known as the *reduced derivatives* of f; they are a by-product of the Taylor-series method for integrating differential equations (Chapter 9, §§ 4–6). As many of them as are significant in (15) are tabulated side by side with f. The process of interpolation is the evaluation of the right-hand side of (15) for the value of p in question, and can be performed either with the use of a table of powers or by treating the curtailed series as a polynomial in p and building it up on a desk machine by nested multiplication (Chapter 6, § 1).

These aids occupy considerable space in a printed table and are not often given.

Economized polynomials

12. In this method the tabulated function $f(x)$ is represented in the interval $a \leqslant x \leqslant a+h$ by a polynomial of the form

$$f(a+ph) = f(a) + c_1 p + c_2 p^2 + c_3 p^3 + \ldots + c_n p^n, \tag{16}$$

the degree of which is chosen to be as small as possible subject to the condition that the error of the representation shall not exceed half a unit in the last decimal. The coefficients $c_1, c_2, \ldots, c_n$ are tabulated side by side with f, and the interpolation is carried out by direct evaluation of the polynomial on the right of (16).

From the standpoint of the user the method resembles the use of reduced derivatives; the essential difference is that the latter requires more terms.

The coefficients $c_1, c_2, \ldots, c_n$ may be determined from the expansion of $f(a+ph)$ in Chebyshev polynomials (Chapter 8) for the range $0 \leqslant p \leqslant 1$. In this way expansions for the coefficients can be derived in series of central differences similar to (12); see **[151]**.

13. The effect of rounding errors in the use of (16) may be minimized by the following device, due to D. B. Gillies. The function values and the coefficients are evaluated with guarding figures. Then $f, c_1, c_2, \ldots$ are rounded in succession in such a way that, with primes denoting the values to be tabulated, $f', f'+c'_1, f'+c'_1+c'_2, \ldots$ are the correctly rounded values of $f, f+c_1, f+c_1+c_2, \ldots$, respectively. The consequent error in an unrounded interpolate cannot then exceed one half. For

$$f = f' + \epsilon_0, \quad f + c_1 = f' + c'_1 + \epsilon_1, \quad f + c_1 + c_2 = f' + c'_1 + c'_2 + \epsilon_2, \quad \ldots, \tag{17}$$

where $|\epsilon_s| \leqslant \frac{1}{2}$. Hence

$$c_s = c'_s + \epsilon_s - \epsilon_{s-1}. \tag{18}$$

The error in an interpolate is accordingly

$$\begin{aligned} &|\epsilon_0 + (\epsilon_1 - \epsilon_0)p + \ldots + (\epsilon_n - \epsilon_{n-1})p^n| \\ &\quad = |\epsilon_0(1-p) + \epsilon_1(p-p^2) + \ldots + \epsilon_{n-1}(p^{n-1} - p^n) + \epsilon_n p^n| \\ &\quad \leqslant \tfrac{1}{2}\{(1-p) + (p-p^2) + \ldots + (p^{n-1} - p^n) + p^n\} = \tfrac{1}{2}. \end{aligned}$$

14. The advantage of this method is the speed and convenience with which an interpolation may be carried out. For the *same* interval, the coefficients c_s require more space than the modified differences, but it is often worthwhile to *increase* the interval and use a polynomial of higher degree.

15. Another form of economized interpolation polynomial is obtained by rearranging (11) in the form

$$f_p = qf_0 + q(1-q^2)\,d_{2,0} + q^3(1-q^2)\,d_{4,0} + pf_1 + p(1-p^2)\,d_{2,1} + p^3(1-p^2)\,d_{4,1}, \tag{19}$$

where $q = 1-p$ and

$$d_2 = \tfrac{1}{6}(-\delta_m^2 + 16\gamma^4), \qquad d_4 = -\tfrac{25}{3}\gamma^4. \tag{20}$$

This has the same type of symmetry as the Everett formulae (6) and (11), but has the advantage that it can be evaluated easily without the aid of tables of interpolation coefficients.

Using a device similar to that of § 13, we can ensure that the accumulated error in an unrounded interpolate never exceeds one half. This compares favourably with the maximum error of 1·1 units associated with (11); see [**150**].

Examples of tables which have aids for use with (16) and (19) are given in [**215**].

Lagrange's method

16. Here the interpolate f_p is computed from a number of consecutive tabular values surrounding the desired value of the argument. With an odd number $(2n+1)$ of points Lagrange's formula may be written as

$$f_p = L_{-n}(p)f_{-n} + L_{-n+1}(p)f_{-n+1} + \dots + L_n(p)f_n, \tag{21}$$

and with an even number $(2n)$ as

$$f_p = L_{-n+1}(p)f_{-n+1} + L_{-n+2}(p)f_{-n+2} + \dots + L_n(p)f_n. \tag{22}$$

The advantages of this method are wholly on the side of the compiler; no interpolation aids are given. From the standpoint of the user there are several drawbacks. The calculation is laborious; bulky tables of the Lagrange coefficients $L_s(p)$ are required; there is difficulty in deciding how many points to use, particularly if the full accuracy of the table is not required; and interpolation near the ends of a table may be troublesome.

It is true that by omitting interpolation aids, space is made available which could be filled with additional function values, thus permitting a reduction of the interval. This would have the effect of reducing the degree of the Lagrange polynomial needed for interpolation. Nevertheless, the more powerful aids at the unreduced interval are almost invariably faster to use than the Lagrange formula at the smaller interval.

USE OF AUXILIARY VARIABLES

17. We have assumed, in effect, that for the purpose of interpolation the tabulated function can be represented reasonably as a polynomial. It may not always be possible to do this directly and the compiler must then introduce new variables, dependent or independent or even both. We mention here some of the circumstances in which this need may arise.

Singularities

18. Suppose, for example, that near $x = 0$ we have

$$f(x) = x^{-1} + \phi(x), \tag{23}$$

where $\phi(x)$ is a well-behaved function. Clearly $f(x)$ cannot be interpolated directly by a polynomial near $x = 0$, but we may tabulate either $f(x) - x^{-1}$ or $xf(x)$, both of which are well behaved there.

A common form of singularity involves the logarithmic function. For example, near $x = 0$ the exponential integral

$$-\mathrm{Ei}(-x) = \int_x^\infty \frac{e^{-t}}{t}\,dt \tag{24}$$

has the series expansion

$$-\mathrm{Ei}(-x) = -\gamma - \ln x + \sum_{1}^{\infty} (-)^{n-1} \frac{x^n}{n.n!} \quad (x > 0), \qquad (25)$$

where $\gamma = 0{\cdot}577\ldots$ is Euler's constant. Accordingly, for interpolation purposes, near $x = 0$ we tabulate $-\mathrm{Ei}(-x) + \ln x$ rather than $-\mathrm{Ei}(-x)$ itself.

Singularities at infinity

19. In order to make a table which can be used for indefinitely large values of the argument x, we must take a new argument such as x^{-1}.

For example, for large x the exponential integral has the asymptotic expansion

$$-\mathrm{Ei}(-x) \sim \frac{e^{-x}}{x}\left(1 - \frac{1!}{x} + \frac{2!}{x^2} - \ldots\right) = \frac{e^{-x}}{x} S(x), \text{ say.} \qquad (26)$$

With argument $z \equiv x^{-1}$ the function S can be computed for values of z between $0{\cdot}0$ and $0{\cdot}1$, say, and it is easily interpolable to about seven decimals at the interval $0{\cdot}01$ in z. The required function is then obtainable by multiplication by e^{-x}/x, readily determined from exponential tables.

20. As a second example, for large x the Bessel functions $J_0(x)$ and $Y_0(x)$ are oscillatory functions with period approximately equal to 2π. This can be seen from the asymptotic expansions

$$J_0(x) = \left(\frac{2}{\pi x}\right)^{\frac{1}{2}} (P\cos\theta - Q\sin\theta), \quad Y_0(x) = \left(\frac{2}{\pi x}\right)^{\frac{1}{2}} (P\sin\theta + Q\cos\theta), \qquad (27)$$

where $\theta = x - \frac{1}{4}\pi$ and

$$P \sim 1 - \frac{1^2.3^2}{2!\,(8x)^2} + \frac{1^2.3^2.5^2.7^2}{4!\,(8x)^4} - \ldots, \quad Q \sim -\frac{1^2}{1!\,(8x)} + \frac{1^2.3^2.5^2}{3!\,(8x)^3} - \ldots. \qquad (28)$$

For large x we can tabulate P and Q as functions of x^{-1}. Then $J_0(x)$ and $Y_0(x)$ can be found with the aid of tables of the trigonometric functions.

Another pair of auxiliary functions which vary slowly and are well behaved at $x^{-1} = 0$ is $(\frac{1}{2}\pi x)^{\frac{1}{2}} M$ and $\psi - x$, where M and ψ are defined by

$$J_0(x) = M\cos\psi, \qquad Y_0(x) = M\sin\psi. \qquad (29)$$

M and ψ are sometimes called the *modulus function* and *phase function* respectively.

21. Besides being necessary for satisfactory interpolation, auxiliary functions are often easier to compute than the original functions.

For example, if we substitute

$$y = Me^{i\psi} \qquad (30)$$

in the differential equation (4) satisfied by $J_0(x)$ and $Y_0(x)$, divide throughout by $e^{i\psi}$ and separate real and imaginary parts, we obtain

$$M'' - M\psi'^2 + x^{-1}M' + M = 0, \qquad M\psi'' + 2M'\psi' + x^{-1}M\psi' = 0. \qquad (31)$$

The second of these equations can be re-expressed as

$$2\frac{M'}{M} + \frac{\psi''}{\psi'} + \frac{1}{x} = 0, \qquad (32)$$

and then integrated immediately to give

$$M^2\psi' = cx^{-1}, \tag{33}$$

where c is a constant. Substitution of (33) in the first of (31) now gives a differential equation for M:

$$M'' + \frac{1}{x}M' + M - \frac{c^2}{x^2 M^3} = 0. \tag{34}$$

The value of c depends on the normalization of the solutions; for the choice (29) we have $c = 2/\pi$.

The equation (34) can be integrated numerically for M, for example by the method of § 21 of Chapter 9, and ψ subsequently evaluated by quadrature of $c/(xM^2)$. Although non-linear, equation (34) is easier to integrate than the original equation (4), because the rapid oscillations present in the solutions of (4) have effectively been removed. Accordingly, the integration can be carried out at a much larger interval.

This procedure is of general applicability to modulus and phase functions associated with oscillatory solutions of linear second-order differential equations.

TABLES FOR AUTOMATIC COMPUTERS

22. The conventional type of mathematical table is inconvenient to use with an automatic computer because of the excessive amount of storage required. The storage requirement can be reduced only at the expense of having an elaborate interpolation routine.

In the case of elementary functions simple properties can often be used to evaluate them directly, and the need for tables avoided. The processes used include iteration, recurrence, summation of power series, evaluation of continued fractions and even the solution of differential equations.

23. It is not always possible or convenient to use these methods. A general method is to represent the wanted function $f(x)$ over a large range $a \leqslant x \leqslant b$ by an expansion in Chebyshev series

$$f(x) = \tfrac{1}{2}a_0 + a_1 T_1(t) + a_2 T_2(t) + \dots, \tag{35}$$

where

$$t = \frac{2x-a-b}{b-a}. \tag{36}$$

The values of the coefficients $a_0, a_1, a_2, \dots$ then comprise a compact *machine table* from which $f(x)$ can be evaluated with the aid of a subroutine based on the algorithm given in Chapter 8, § 16. Machine tables of this kind for elementary functions and certain higher functions of a single variable are given in **[165]** and **[214]**; Table 1 of Chapter 8 provides a typical example.

24. In planning future mathematical tables first consideration should be given to a machine form. This form can then be used to generate the orthodox table, and might often, with advantage, be published together with it. Indeed, as more automatic computers become available, it is likely that for many functions machine tables will supplant the conventional form. This is particularly true of functions of several variables, orthodox tables of which generally do not fulfil interpolation requirements satisfactorily.

BIBLIOGRAPHY

INTRODUCTION

It should be emphasized at the outset that facility in computation can only be gained by *practice*; reading alone is not sufficient. This is true of both desk-machine and automatic work. In the latter case moreover, programmers must be acquainted with the fundamentals of good desk-machine practice before they can be efficient and reliable.

The books and papers given in this Bibliography are mainly in the English language and have been selected to provide amplification of the subject matter of this manual, and suitable introductions to more advanced work. The list is divided into sections corresponding to the subjects of the previous chapters. In addition, there are sections at the end on tables, facts and formulae, curve-fitting and smoothing, harmonic analysis, integral equations, miscellanea, computing machines, nomography, and journals and reviews. There is, of course, considerable overlap between some of the sections; for example, the theory of finite-difference processes and the theory of linear equations and matrices are both required in the solution of ordinary and partial differential equations and integral equations.

GENERAL

1. Buckingham, R. A. 1957 *Numerical methods*. London: Pitman.

2. Hartree, D. R. 1958 *Numerical analysis*. Second edition (first edition 1952). Oxford University Press.

3. Kunz, K. S. 1957 *Numerical analysis*. New York and London: McGraw-Hill.

Although intended primarily for desk-machine users, these three books all serve as excellent introductions to the practical aspects of modern computation. Each is clearly written and contains a wealth of information.

Of the three, the longest and most complete treatment is that of Buckingham. Almost all branches of numerical analysis are covered, and little of practical importance in the literature has escaped the author's attention.

Hartree's book is much shorter. The basic processes of numerical calculus are fully treated; the shortening is at the expense of more advanced topics. For example, only twenty pages are allotted to partial differential equations, and none at all to integral equations.

The scope of Kunz's book is similar to that of Buckingham. A defect is the inadequate treatment of latent root (eigenvalue) problems. Other omissions concern some contributions by British workers: 'throwback' interpolation (Comrie and others), interpolation by cross means (Aitken), and implicit methods for parabolic partial differential equations (Crank and Nicolson). The student may perhaps be in greater need of assistance from his supervisor with this book than with the other two.

4. Hildebrand, F. B. 1956 *Introduction to numerical analysis*. New York and London: McGraw-Hill.

Gives an excellent and detailed account of the theory and practice of interpolation, numerical differentiation, numerical integration, and many kinds of approximation methods, including the use of orthogonal polynomials. The numerical solution of ordinary differential equations, polynomial equations, and simultaneous algebraic equations is deliberately treated in less detail. This book is intended for the serious student of numerical analysis rather than the occasional computer.

5. Kopal, Z. 1955 *Numerical analysis*. New York and London: John Wiley.

Concentrates on the application of numerical techniques to problems of infinitesimal calculus in a single variable. Covers much of the same ground as Hildebrand in interpolation, differentiation and integration, with extra material on ordinary differential equations, including boundary-value problems, and integral equations. There is, deliberately, nothing on the solution of algebraic or transcendental equations. 'Throwback' interpolation is discussed in some detail, and there is an interesting historical introduction.

6. Lanczos, C. 1957 *Applied analysis*. New York: Prentice-Hall; London: Pitman.

A very readable book on practical computing methods, with good chapters on polynomial equations, matrices and latent root problems, data analysis, and quadrature, but including no systematic account of methods for the solution of differential and integral equations. Some of the material is not readily available elsewhere, particularly the chapters on harmonic analysis and Chebyshev series, but much standard material is not included.

7. Householder, A. S. 1953 *Principles of numerical analysis*. New York and London: McGraw-Hill.

Contains chapters on error analysis, solution of linear equations, inversion of matrices and the determination of their latent roots and vectors, solution of polynomial equations, interpolation, and numerical differentiation and integration. This is a mathematical text-book rather than an introduction to computing practices.

8. Alt, F. L. 1958 *Electronic digital computers*. New York and London: Academic Press.

This appears to be the first substantial general treatment of the practical problems of numerical analysis from the standpoint of automatic computation, almost half the book being devoted to this subject. It covers linear equations, latent roots, differentiation, quadrature, polynomial and transcendental equations, and ordinary and partial differential equations. See also [**209**] and [**210**].

9. Scarborough, J. B. 1958 *Numerical mathematical analysis*. Fourth edition (first edition 1930). Baltimore: The Johns Hopkins Press; Oxford University Press.

10. Milne, W. E. 1949 *Numerical calculus*. Princeton University Press.

11. Whittaker, E. T. and Robinson, G. 1944 *The calculus of observations*. Fourth edition (first edition 1924). London: Blackie.

These three works are among the older text-books. They contain useful chapters on finite-difference theory, interpolation, differentiation, quadrature, processing of observational data, and, in the case of the first two, simultaneous linear equations. Zeros of polynomials and the solution of differential equations are also treated, but rather incompletely.

12. Forsythe, G. E. and Rosenbloom, P. C. 1958 *Numerical analysis and partial differential equations*. New York and London: John Wiley.

The title is misleading. The book consists of two parts: the first, by Forsythe, is a brief survey of the contemporary state of numerical analysis; the second and greater part, by Rosenbloom, is concerned with existence theory in abstract spaces of solutions of partial differential equations. Forsythe's survey includes information not readily available elsewhere on Russian contributions to numerical analysis.

LINEAR EQUATIONS, MATRICES, LATENT ROOTS AND VECTORS (CHAPTERS 1–5)

A concise account of the elementary mathematical properties of matrices is given in the following monograph.

13. Aitken, A. C. 1956 *Determinants and matrices*. Ninth edition (first edition 1939). Edinburgh: Oliver and Boyd; New York: Interscience.

From the more advanced mathematical treatises, the following book and [**212**] may be selected for meeting the requirements of the modern numerical analyst.

14. Bellman, R. 1960 *Introduction to matrix analysis*. New York and London: McGraw-Hill.

Solution of linear equations and inversion of matrices

Suitable introductions are included in [**1**], [**2**], [**3**] and in the following book.

15. Crandall, S. H. 1956 *Engineering analysis*. New York and London: McGraw-Hill.

16. Fox, L. 1954 Practical solution of linear equations and inversion of matrices. *Appl. Math. Ser. U.S. Bur. Stand.* **39**, 1–54. Washington: Government Printing Office.

A comprehensive account of modern methods from the standpoint of the desk-machine user. See also [**212**].

17. Fox, L. and Hayes, J. G. 1951 More practical methods for the inversion of matrices. *J. R. Statist. Soc.* B, **13**, 83–91.

Describes, with full practical details, some good desk-machine methods for inverting matrices.

18. Householder, A. S. 1957 A survey of some closed methods for inverting matrices. *J. Soc. Indust. Appl. Math.* **5**, 155–169.

In this more advanced paper the equivalence of various direct methods is investigated.

Large matrices having a high proportion of zero elements occur in the solution of elliptic partial differential equations by finite-difference methods. The indirect solution of the corresponding linear equations is the subject of [**112**] to [**116**].

Latent roots (eigenvalues) and latent vectors (eigenvectors)

Introductions to the practical side are given in [**1**] and [**212**], and a comprehensive treatment in [**216**]. See also [**7**].

19. Wilkinson, J. H. 1954 The calculation of the latent roots and vectors of matrices on the pilot model of the ACE. *Proc. Camb. Phil. Soc.* **50**, 536–566.

A comprehensive and detailed account of the practical aspects of iteration and matrix powering, written from the standpoint of automatic computation. The treatment of complex roots and root-removal processes is particularly thorough.

20. White, P. A. 1958 The computation of eigenvalues and eigenvectors of a matrix. *J. Soc. Indust. Appl. Math.* **6**, 393–437.

A useful survey.

21. Givens, W. 1954 Numerical computation of the characteristic values of a real symmetric matrix. Oak Ridge National Laboratory, ORNL-1574.

This rather inaccessible reference appears to be the only one giving a full account of the rotations method described in Chapter 3, § 16. An extensive error analysis of the process is included.

22. Wilkinson, J. H. 1958 The calculation of the eigenvectors of codiagonal matrices. *Computer J.* 1, 90–96.

Supplements the method of the preceding paper for computing latent roots, by a practical procedure for computing accurately the latent vectors.

23. Givens, W. 1958 Computation of plane unitary rotations transforming a general matrix to triangular form. *J. Soc. Indust. Appl. Math.* **6**, 26–50.

The method described in [**21**] is applied to unsymmetric matrices, reducing them to almost triangular (Hessenberg) form. A further transformation is then applied to achieve reduction to triple-diagonal form, but this is numerically unstable.

24. Wilkinson, J. H. 1959 Stability of the reduction of a matrix to almost triangular and triangular forms by elementary similarity transformations. *J. Ass. Comp. Mach.* **6**, 336–359.

Shows how to overcome the numerical instability associated with the method of the preceding reference.

25. Wilkinson, J. H. 1960 Householder's method for the solution of the algebraic eigenproblem. *Computer J.* **3**, 23–27.

Describes the practical application of a powerful method for transforming a symmetric matrix to triple-diagonal form by orthogonal transformations. This method, which has advantages over Givens' method, was first described in [**28**] below.

26. Lanczos, C. 1950 An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.* **45**, 255–282.

Describes the so-called 'method of minimized iterations', which is another direct method for the transformation of a matrix, symmetric or unsymmetric, into triple–diagonal form.

27. Wilkinson, J. H. 1958 The calculation of eigenvectors by the method of Lanczos. *Computer J.* **1**, 148–152.

Supplements the preceding paper by showing how to overcome practical difficulties in the accurate computation of the latent vectors.

28. Householder, A. S. and Bauer, F. L. 1959 On certain methods for expanding the characteristic polynomial. *Numerische Math.* **1**, 29–37.

A survey of methods for reducing matrices to Hessenberg form by similarity transformations, that is, transformations of the form $\mathbf{H}^{-1}\mathbf{A}\mathbf{H}$.

29. Rutishauser, H. 1958 Solution of eigenvalue problems with the LR-transformation. *Appl. Math. Ser. U.S. Bur. Stand.* **49**, 47–81.

This is a recently discovered iterative method applicable to symmetric and unsymmetric matrices, but particularly useful for band matrices and Hessenberg matrices. It yields the latent roots in order, starting with the smallest.

30. Givens, W. 1953 A method of computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices. *Appl. Math. Ser. U.S. Bur. Stand.* **29**, 117–122.

31. Ortega, J. M. 1960 On Sturm sequences for tridiagonal matrices. Applied Mathematics and Statistics Laboratories, Stanford University, *Tech. Rep.* No. 4.

These two references between them establish the full form of the Sturm-sequence theorem required for the method of bisections (Chapter 3, § 18).

Error analysis

See [**19**] and [**21**].

32. Wilkinson, J. H. 1960 Rounding errors in algebraic processes. *Information Processing*, 44–53. Paris: UNESCO; Munich: Oldenbourg; London: Butterworths.

Contains a fuller development of the ideas of Chapter 5 for the solution of linear algebraic equations using fixed-point arithmetic. Its new approach is the most powerful one available for assessing errors arising in practical matrix problems. A book on the subject by the same author is in course of preparation. See also [**213**].

33. Householder, A. S. 1958 The approximate solution of matrix problems. *J. Ass. Comp. Mach.* **5**, 205–243.

Defines matrix norms, investigates their properties and applies the results to error assessment. A more elementary introduction to matrix norms is given in [**212**].

34. Turing, A. M. 1948 Rounding-off errors in matrix processes. *Quart. J. Mech.* **1**, 287–308.

35. von Neumann, J. and Goldstine, H. H. 1947 Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.* **53**, 1021–1099.

These two earlier references derive theoretical upper bounds for the errors.

ZEROS OF POLYNOMIALS (CHAPTER 6)

Suitable introductions are included in [**1**] to [**4**]; [**1**] is particularly good.

36. Wilkinson, J. H. 1959 The evaluation of the zeros of ill-conditioned polynomials. *Numerische Math.* **1**, 150–180.

A complete account of modern automatic procedures. It includes an assessment of the ill-conditioned nature of the problem and of iterative techniques.

37. Olver, F. W. J. 1952 The evaluation of zeros of high-degree polynomials. *Phil. Trans.* A, **244**, 385–415.

A survey from the standpoint of desk computation, including iterative methods, root-squaring, and the Aitken–Bernoulli method.

38. Muller, D. E. 1956 A method for solving algebraic equations using an automatic computer. *Math. Tab., Wash.* **10**, 208–215.

Describes a powerful iterative process which is often used in automatic work.

39. Aitken, A. C. 1926 On Bernoulli's numerical solution of algebraic equations. *Proc. Roy. Soc. Edinb.* **46**, 289–305.

Gives an elegant generalization of Bernoulli's method, though this method is seldom used in practice. Also includes an algorithm, the 'δ^2-process' (Chapter 13, § 3), for accelerating the convergence of sequences.

Equations of low degree

40. Porter, A. and Mack, C. 1949 New methods for the numerical solution of algebraic equations. *Phil. Mag.* **40**, 578–585.

Describes a method of successive approximation applicable when the degree does not exceed six. Also refers to a war-time RRDE report which gives charts (nomograms) for the solution of cubics and quartics. Charts for the solution of cubics and quartics are also given in [**160**].

41. Salzer, H. E., Richards, C. H. and Arsham, I. 1958 *Table for the solution of cubic equations.* New York and London: McGraw-Hill.

The most comprehensive table of its kind.

THE CALCULUS OF FINITE DIFFERENCES (CHAPTER 7)

See [**1**] to [**5**]. Sound accounts of the practical processes of interpolation, quadrature and numerical differentiation are given in [**167**], [**159**], and in a chapter of the following book.

42. Jeffreys, H. and Jeffreys, B. S. 1956 *Methods of mathematical physics.* Third edition (first edition 1946). Cambridge University Press.

43. Freeman, H. 1960 *Finite differences for actuarial students.* Second edition (first edition 1939, published as *Mathematics for actuarial students, Part II*). Cambridge University Press.

An elementary and readable book.

44. Jordan, C. 1950 *Calculus of finite differences.* Second edition (first edition 1939). New York: Chelsea.

A very full treatment which also includes a chapter on elementary statistics.

45. Mineur, H. 1952 *Techniques de calcul numérique*. Paris and Liége: Librairie Polytechnique Ch. Béranger.

Includes a detailed treatment of interpolation, numerical differentiation and quadrature, with much information on the handling of singularities. Other material on zeros of polynomials and ordinary differential equations is somewhat out of date.

46. Kuntzmann, J. 1959 *Méthodes numériques*. Paris: Dunod.

Entirely devoted to interpolation and numerical differentiation; a very detailed treatment.

47. Bickley, W. G. 1948 Difference and associated operators, with some applications. *J. Math. Phys.* **27**, 183–192.

48. Michel, J. G. L. 1946 Central-difference formulae obtained by means of operator expansions. *J. Inst. Actu.* **72**, 470–480.

These two papers give clear expositions of the operational method for deriving finite-difference formulae.

49. Bickley, W. G. 1939 Formulae for numerical integration. *Math. Gaz.* **23**, 352–359.

An extensive collection of finite-difference quadrature formulae. There is a similar collection for differentiation (*Math. Gaz.* **25**, 19–26, 1941).

50. Sheppard, W. F. 1906 On the accuracy of interpolation by finite differences. *Proc. Lond. Math. Soc.* **4**, 320–341.

This examines the rounding errors associated with the use of various finite-difference formulae.

51. Milne-Thomson, L.M. 1951 *The calculus of finite differences*. Reprint (first edition 1933). London: Macmillan.

52. Steffensen, J. F. 1950 *Interpolation*. Second edition (first English edition 1927). New York: Chelsea.

53. Nörlund, N. E. 1924 *Vorlesungen über Differenzenrechnung*. Berlin: Springer.

These three excellent treatises discuss the subject from a more theoretical and mathematically rigorous standpoint.

CHEBYSHEV SERIES (CHAPTER 8)

See [**6**] and [**214**]. Other readable expositions are contained in

54. Appl. Math. Ser. U.S. Bur. Stand. **9**, 1952 *Tables of Chebyshev polynomials*. Washington: Government Printing Office.

and

55. Lanczos, C. 1938 Trigonometric interpolation of empirical and analytical functions. *J. Math. Phys.* **17**, 123–199.

56. Bernstein, S. 1926 *Leçons sur les propriétés extrémales des fonctions analytiques*. Paris: Gauthier-Villars.

A treatise on the fundamental problems of approximation by polynomials.

57. Clenshaw, C. W. 1955 A note on the summation of Chebyshev series. *Math. Tab., Wash.* **9**, 118–120.

Derives an algorithm for evaluating the sum of the series, and investigates the possibility of error build-up in applications.

58. Hornecker, G. 1958 Evaluation approchée de la meilleure approximation polynomiale d'ordre n de $f(x)$ sur un segment fini [a,b]. *Chiffres* **1**, 157–169.

The coefficients of the polynomial of 'best fit', that is, of least maximum deviation from $f(x)$, are expressed approximately in terms of the Chebyshev coefficients.

59. Murnaghan, F. D. and Wrench, J. W. 1959 The determination of the Chebyshev approximating polynomial for a differentiable function. *Math. Tab., Wash.* **13**, 185–193.

This is representative of a number of recent papers giving iterative procedures for computing the polynomial of best fit.

ORDINARY DIFFERENTIAL EQUATIONS (CHAPTERS 9, 10)

General

See [**1**], [**2**], [**3**].

60. Collatz, L. 1960 *The numerical treatment of differential equations.* (Translated by P. G. Williams.) Third edition (first edition 1951). Berlin: Springer.

Probably the most thorough single account of solving both ordinary and partial differential equations of all kinds.

61. Fox, L. 1954 A note on the numerical integration of first-order differential equations. *Quart. J. Mech.* **7**, 367–378.

Examines general aspects of the numerical solution of first-order equations, linear or otherwise.

Taylor-series method

See [**1**] to [**5**].

62. Wilson, E. M. 1949 A note on the numerical integration of differential equations. *Quart. J. Mech.* **2**, 208–211.

Gives a refinement of the method applicable in certain cases.

Predictor-corrector methods

See [**3**], [**4**], [**10**].

63. Milne, W. E. 1953 *Numerical solution of differential equations.* New York and London: John Wiley.

64. Hamming, R. W. 1959 Stable predictor-corrector methods for ordinary differential equations. *J. Ass. Comp. Mach.* **6**, 37–47.

Central-difference methods

See [**2**], [**42**], [**167**].

Deferred-correction methods

65. Fox, L. and Goodwin, E. T. 1949 Some new methods for the numerical integration of ordinary differential equations. *Proc. Camb. Phil. Soc.* **45**, 373–388.

66. Clenshaw, C. W. and Olver, F. W. J. 1951 Solution of differential equations by recurrence relations. *Math. Tab., Wash.* **5**, 34–39.

See also [**1**].

Runge-Kutta methods

See [**3**], [**4**], [**5**], [**60**].

67. Kutta, W. 1901 Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. *Z. Math. Phys.* **46**, 435–453.

Includes many formulae.

68. Gill, S. 1951 A process for the step-by-step integration of differential equations in an automatic digital computing machine. *Proc. Camb. Phil. Soc.* **47**, 96–108.

Describes an efficient modification in which the intermediate points are chosen to minimize the storage requirements and the effect of rounding error.

69. Martin, D. W. 1958 Runge–Kutta methods for integrating differential equations on high-speed digital computers. *Computer J.* **1**, 118–123.

Describes further modifications with a view to a gain in speed.

70. de Vogelaere, R. 1955 A method for the numerical integration of differential equations of second order without explicit first derivatives. *J. Res. Nat. Bur. Stand.* **54**, 119–125.

A hybrid method useful for equations of the form $y'' = f(x, y)$.

Chebyshev-series method

The 'τ-method' of Lanczos is described in [**6**] and the introduction to [**54**].

71. Clenshaw, C. W. 1957 The numerical solution of linear differential equations in Chebyshev series. *Proc. Camb. Phil. Soc.* **53**, 134–149.

The method described and illustrated by examples in this paper has advantages over the τ-method.

Building-up errors; stability

See [**4**], [**5**], [**60**], [**64**], [**65**].

72. Sterne, T. E. 1953 The accuracy of numerical solutions of ordinary differential equations. *Math. Tab., Wash.* **7**, 159–164.

73. Dahlquist, G. 1956 Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.* **4**, 33–53.

74. Dahlquist, G. 1959 Stability and error bounds in the numerical integration of ordinary differential equations. *K. Tekn. Högsk. Handl.* **130**.

75. Carr, J. W. 1958 Error bounds for the Runge–Kutta single-step integration process. *J. Ass. Comp. Mach.* **5**, 39–44.

Boundary-value problems

See [**60**], [**71**].

76. Fox, L. 1957 *The numerical solution of two-point boundary problems in ordinary differential equations.* Oxford University Press.

This comprehensive work covers direct matrix methods, relaxation, use of initial-value techniques, and eigenvalue problems. Numerous examples are given.

77. Fox, L. 1949 The solution by relaxation methods of ordinary differential equations. *Proc. Camb. Phil. Soc.* **45**, 50–68.

Relaxation methods are also described in [**105**].

The application of initial-value techniques to boundary-value problems is discussed in [**1**] to [**5**], [**63**], and in the following paper.

78. Warner, F. J. 1957 On the solution of 'jury' problems with many degrees of freedom. *Math. Tab., Wash.* **11**, 268–271.

The computation of eigenvalues is treated in many of the foregoing references. See also the following treatise:

79. Collatz, L. 1945 *Eigenwertprobleme und ihre numerische Behandlung.* Leipzig: Becker and Erler.

PARTIAL DIFFERENTIAL EQUATIONS (CHAPTERS 11, 12)

Short introductions are given in [**1**], [**2**], [**3**]. For more extensive treatments see [**60**] and [**217**].

80. Lowan, A. N. 1957 *The operator approach to problems of stability and convergence of solutions of difference equations and the convergence of various iteration*

procedures. New York: Scripta Mathematica; Washington: Office of Technical Services.

A connected account of the solution of stability and convergence problems by use of matrix theory. All three kinds of partial difference equation are treated.

Hyperbolic equations

See [**15**], [**60**] and [**210**].

81. Hartree, D. R. 1953 Some practical methods of using characteristics in the calculation of non-steady compressible flow. Harvard University, Dept. of Maths. *Rep. No.* LA-HU-1.

A good readable introduction, though rather inaccessible.

82. Courant, R. and Hilbert, D. 1937 *Methoden der Mathematischen Physik* II. Berlin: Springer.

83. Courant, R. and Friedrichs, K. O. 1948 *Supersonic flow and shock waves*. New York and London: Interscience.

These two references give information on the physics of common problems, as well as on numerical methods of solution.

84. Courant, R., Isaacson, E. and Rees, M. 1952 On the solution of nonlinear hyperbolic differential equations by finite differences. *Commun. Pure Appl. Math.* **5**, 243–255.

See also [**89**].

85. Douglas, J. 1956 On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations. *J. Soc. Indust. Appl. Math.* **4**, 20–37.

This aspect is also treated in [**80**].

86. von Neumann, J. and Richtmyer, R. D. 1950 A method for the numerical calculation of hydrodynamic shocks. *J. Appl. Phys.* **21**, 232–237.

87. Fox, P. and Ralston, A. 1956 On the numerical solution of the equations for spherical waves of finite amplitude, I. *J. Math. Phys.* **35**, 313–328.

88. Roberts, L. 1956 On the numerical solution of the equations for spherical waves of finite amplitude, II. *J. Math. Phys.* **35**, 329–337.

These three papers examine difficulties in carrying out numerical work in the neighbourhood of a shock wave.

Parabolic equations

See [**60**].

89. Richtmyer, R. D. 1957 *Difference methods for initial-value problems*. New York and London: Interscience.

90. Carslaw, H. S. and Jaeger, J. C. 1959 *Conduction of heat in solids*. Second edition (first edition 1946). Oxford University Press.

A standard work of reference on analytical solutions of the heat-conduction equation. Numerical methods are treated very briefly in the final chapter.

91. Crank, J. 1956 *The mathematics of diffusion*. Oxford University Press.

A text-book on analytical and numerical methods of solving diffusion problems.

92. Eyres, N. R., Hartree, D. R., Ingham, J., Jackson, R., Sarjant, R. J. and Wagstaff, J. B. 1946 The calculation of variable heat flow in solids. *Phil. Trans.* A, **240**, 1–57.

93. Crank, J. and Nicolson, P. 1947 A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. Phil. Soc.* **43**, 50–67.

The original reference to the Crank–Nicolson method. Comparison is made with methods given in the preceding reference.

94. Todd, J. 1956 A direct approach to the problem of stability in the numerical solution of partial differential equations. *Commun. Pure Appl. Math.* **9**, 597–612.

Stability is investigated by matrix theory. The paper contains an extensive bibliography. See also [**80**].

95. O'Brien, G. G., Hyman, M. A. and Kaplan, S. 1951 A study of the numerical solution of partial differential equations. *J. Math. Phys.* **29**, 223–251.

Investigates stability by the method of von Neumann. See also [**85**].

96. Albasiny, E. L. 1960 On the numerical solution of a cylindrical heat-conduction problem. *Quart. J. Mech.* **13**, 374–384.

Includes an investigation of the effect of neglecting singularities in the boundary conditions in setting up the finite-difference equations. Singularities are also treated in [**91**] and the following reference.

97. Crank, J. 1957 Two methods for the numerical solution of moving-boundary problems in diffusion and heat flow. *Quart. J. Mech.* **10**, 220–231.

98. Conte, S. D. 1957 A stable implicit finite-difference approximation to a fourth-order parabolic equation. *J. Ass. Comp. Mach.* **4**, 18–23.

99. Peaceman, D. W. and Rachford, H. H. 1955 The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.* **3**, 28–41.

Gives a modification of the implicit method for solving the heat-conduction equation with two space variables. The method can be used to find the steady-state solution, hence solving Laplace's equation; see also [**116**].

100. Douglas, J. 1955 On the numerical integration of $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial u}{\partial t}$ by implicit methods. *J. Soc. Indust. Appl. Math.* **3**, 42–65.

Elliptic equations: relaxation methods for desk machines

See [**3**], [**60**].

101. Bickley, W. G. 1948 Finite-difference formulae for the square lattice. *Quart. J. Mech.* **1**, 35–42.

102. Shaw, F. S. 1958 *An introduction to relaxation methods.* Second edition (first edition 1953). New York: Dover; London: Constable.

103. Allen, D. N. de G. 1954 *Relaxation methods.* New York and London: McGraw-Hill.

These two books teach the basic computational processes of the relaxation method.

104. Southwell, R. V. 1946 and 1956 *Relaxation methods in theoretical physics.* Volume 1 (1946), Volume 2 (1956). Oxford University Press.

Contains examples of the application of the relaxation process to physical problems; Volume 1 covers second-order equations and Volume 2 fourth-order equations, in two independent variables. See also [**15**].

105. Fox, L. 1947 Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations. *Proc. Roy. Soc.* A, **190**, 31–59.

Gives a full treatment of the 'difference-correction' technique.

106. Fox, L. 1950 The numerical solution of elliptic differential equations when the boundary conditions involve a derivative. *Phil. Trans.* A, **242**, 345–378.

107. Viswanathan, R. V. 1957 Solution of Poisson's equation by relaxation method—normal gradient specified on curved boundaries. *Math. Tab., Wash.* **11**, 67–78.

108. Motz, H. 1946 The treatment of singularities of partial differential equations by relaxation methods. *Quart. Appl. Math.* **4**, 371–377.

109. Woods, L. C. 1953 The relaxation treatment of singular points in Poisson's equation. *Quart. J. Mech.* **6**, 163–185.

Singularities are also treated in [**42**].

Elliptic equations: methods for automatic computers

110. Karlqvist, O. 1952 Numerical solution of elliptic difference equations by matrix methods. *Tellus* **4**, 374–384.

111. Cornock, A. F. 1954 The numerical solution of Poisson's and the bi-harmonic equations by matrices. *Proc. Camb. Phil. Soc.* **50**, 524–535.

These two papers describe direct methods for solving the algebraic equations which represent the partial differential equation.

112. Engeli, M., Ginsburg, Th., Rutishauser, H. and Stiefel, E. 1959 *Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary-value problems.* Basle: Birkhäuser.

A detailed study of theoretical and practical aspects of iterative methods, including gradient methods and successive overrelaxation, containing valuable basic material.

113. Carré, B. A. 1961 The determination of the optimum accelerating factor for successive overrelaxation. *Computer J.* (*In press.*)

Describes an automatic process for solving equations with Property A (Chapter 4, § 9).

114. Young, D. M. 1955 ORDVAC solutions of the Dirichlet problem. *J. Ass. Comp. Mach.* **2**, 137–161.

An account of practical experience of using the method of successive overrelaxation.

115. Keller, H. B. 1958 On some iterative methods for solving elliptic difference equations. *Quart. Appl. Math.* **16**, 209–226.

Classifies some of the iterative processes and compares their efficiencies.

116. Conte, S. D. and Dames, R. T. 1960 On an alternating direction method for solving the plate problem with mixed boundary conditions. *J. Ass. Comp. Mach.* **7**, 264–273.

Describes the solution of the biharmonic equation by the method of [**99**], and contains references to other applications of this method.

EVALUATION OF LIMITS (CHAPTER 13)

General theory of iterative processes

See [**7**].

117. Hartree, D. R. 1949 Notes on iterative processes. *Proc. Camb. Phil. Soc.* **45**, 230–236.

Defines and classifies the various types of process.

118. Ostrowski, A. M. 1958 A method of speeding up iterations with super-linear convergence. *J. Math. Mech.* **7**, 117–120.

Sums of series; limits of sequences

Brief treatments are given in [**1**], [**2**], [**3**], [**4**].

119. Richardson, L. F. and Gaunt, J. A. 1926 The deferred approach to the limit. *Phil. Trans.* A, **226**, 299–361.

120. Shanks, D. 1955 Non-linear transformations of divergent and slowly convergent sequences. *J. Math. Phys.* **34**, 1–42.

Surveys earlier work and gives a generalization, the $e_m(S_n)$ transformation, of the δ^2-process of Aitken [**39**].

121. Wynn, P. 1956 On a device for computing the $e_m(S_n)$ transformation. *Math. Tab., Wash.* **10**, 91–96.

122. Salzer, H. E. 1954 A simple method for summing certain slowly convergent series. *J. Math. Phys.* **33**, 356–359.

Describes a numerical method applicable when the nth partial sum behaves like a polynomial in $1/n$.

123. Bickley, W. G. and Miller, J. C. P. 1936 The numerical summation of slowly convergent series of positive terms. *Phil. Mag.* **22**, 754–767.

124. Airey, J. R. 1937 The 'converging factor' in asymptotic series and the calculation of Bessel, Laguerre and other functions. *Phil. Mag.* **24**, 521–552.

These two references are among many which describe methods based on particular asymptotic forms of the higher terms of the series. In these two papers expansions are derived in descending powers of n for the *converging factor*, defined as

$$(v_{n+1}+v_{n+2}+\ \dots)/v_n,$$

where v_n is the nth term of the series.

Continued fractions

Elementary properties are given in [**4**] and [**51**].

125. Wall, H. S. 1948 *Analytic theory of continued fractions.* New York: Van Nostrand.

126. Perron, O. 1954 *Die Lehre von den Kettenbrüchen.* Third edition (first edition 1913). Stuttgart: Teubner.

These two references are standard works on the general theory of continued fractions.

127. Teichroew, D. 1952 Use of continued fractions in high-speed computing. *Math. Tab., Wash.* **6**, 127–133.

128. Macon, N. and Baskervill, M. 1956 On the generation of errors in the digital evaluation of continued fractions. *J. Ass. Comp. Mach.* **3**, 199–202.

129. Wynn, P. 1959 Converging factors for continued fractions. *Numerische Math.* **1**, 272–320.

The quotient-difference algorithm

The quotient-difference (QD) algorithm is essentially a procedure for transforming a power series into a continued fraction, but it has applications in other branches of numerical analysis.

130. Rutishauser, H. 1954 Der Quotienten-Differenzen-Algorithmus. *Z. angew. Math. Phys.* **5**, 233–251.

Describes the algorithm and applies it to the evaluation of zeros of polynomials, obtaining a generalization of Bernoulli's method, essentially the same as that given in [**39**].

131. Rutishauser, H. 1954 Anwendungen des Quotienten-Differenzen-Algorithmus. *Z. angew. Math. Phys.* **5**, 496–508.

Derives an alternative form of the algorithm, called the 'progressive' form, and considers applications to the transformation of series into continued fractions and again to the evaluation of zeros of polynomials.

132. Henrici, P. 1958 The quotient-difference algorithm. *Appl. Math. Ser. U.S. Bur. Stand.* **49**, 23–46.

A self-contained introduction to the algorithm, which uses an approach different from that of the two preceding references. Applications are made to latent roots of matrices, zeros of polynomials and expansions in continued fractions.

133. Wynn, P. 1959 A sufficient condition for the instability of the *q-d* algorithm. *Numerische Math.* **1**, 203–207.

EVALUATION OF INTEGRALS (CHAPTER 14)

Finite-difference methods of quadrature are treated in most of the references given in the sections entitled General and The Calculus of Finite Differences. Gauss-type formulae are also discussed in many of these references; in particular, see [4] and [5].

134. Clenshaw, C. W. and Curtis, A. R. 1960 A method for numerical integration on an automatic computer. *Numerische Math.* **2**, 197–205.

Discusses the basic problems of numerical quadrature from the standpoint of automatic computation, and proposes a new method based on the termwise integration of expansions in Chebyshev polynomials.

135. Abramowitz, M. 1954 On the practical evaluation of integrals. *J. Soc. Indust. Appl. Math.* **2**, 20–35.

Describes many methods and artifices, mainly of an analytical character.

136. Goodwin, E. T. and Staton, J. 1948 Table of $\int_0^\infty \frac{e^{-u^2}}{u+x}\,du$. *Quart. J. Mech.* **1**, 319–326.

Describes methods that were used in the evaluation of this integral on desk machines. This example has been used as a 'guinea pig' by some later writers.

137. Appl. Math. Ser. U.S. Bur. Stand. **37** 1954 *Tables of functions and of zeros of functions.* Washington: Government Printing Office.

Includes a 15-decimal table of the zeros of the first 16 Legendre polynomials and the corresponding weight factors for use in Gauss's quadrature formula, and also a similar 12-decimal table for use with the Laguerre–Gauss quadrature formula.

138. Davis, P. and Rabinowitz, P. 1956 Abscissas and weights for Gaussian quadratures of high order. *J. Res. Nat. Bur. Stand.* **56**, 35–37.

Gives 21-decimal values of the zeros of the Legendre polynomials, and the corresponding weights, for degrees 2, 4, 8, 16, 20, 24, 32, 40, 48.

139. Davis, P. and Rabinowitz, P. 1958 Additional abscissas and weights for Gaussian quadratures of high order: values for $n = 64$, 80 and 96. *J. Res. Nat. Bur. Stand.* **60**, 613–614.

140. Salzer, H. E., Zucker, R. and Capuano, R. 1952 Table of the zeros and weight factors of the first twenty Hermite polynomials. *J. Res. Nat. Bur. Stand.* **48**, 111–116.

141. Goodwin, E. T. 1949 The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)\, e^{-x^2}\, dx$. *Proc. Camb. Phil. Soc.* **45**, 241–245.

Shows that many integrals of this type can be evaluated to high accuracy by use of the trapezoidal rule applied with a large interval. An expression for the error of the representation is determined by means of contour integration.

142. Fettis, H. E. 1955 Numerical calculation of certain definite integrals by Poisson's summation formula. *Math. Tab., Wash.* **9**, 85–92.

An alternative treatment of the problem considered in the preceding reference. The results apply to infinite integrals and integrals of periodic functions taken over a whole number of periods.

143. LUKE, Y. L. 1956 Evaluation of an integral arising in numerical integration near a logarithmic singularity *Math. Tab., Wash.* **10**, 14–21.

144. KAPLAN, E. L. 1952 Numerical integration near a singularity. *J. Math. Phys.* **31**, 1–28.

Contains tables for facilitating integration near singularities of the forms $x^{\pm\frac{1}{2}}$ and $\ln x$.

Singular integrands are also discussed in [**4**], [**5**], [**45**] and [**188**].

145. WATSON, G. N. 1944 *Theory of Bessel functions.* Second edition (first edition 1922). Cambridge University Press.

§ 8.3 of this book contains a lemma ('Watson's lemma') which states precise conditions for deriving the asymptotic expansion of an integral of the form $\int_0^\infty e^{-\nu t}\, F(t)\, dt$ by the termwise integration of the expansion of $F(t)$ in ascending powers of t.

146. MILLER, J. C. P. 1960 Numerical quadrature over a rectangular domain in two or more dimensions. Part 1. Quadrature over a square, using up to sixteen equally spaced points. *Math. Computation*, **14**, 13–20.

One of the more practical papers on the evaluation of double integrals.

TABULATION OF MATHEMATICAL FUNCTIONS (CHAPTER 15)

147. MILLER, J. C. P. 1949 The construction of mathematical tables. *Sci. J. R. Coll. Sci.* **20**, 1–11.

This readable pamphlet gives a brief account of the basic principles of table-making.

148. MILLER, J. C. P. 1950 Checking by differences—I. *Math. Tab., Wash.* **4**, 3–11.

Examines the chance of a difference of a given order, which is entirely composed of rounding errors, exceeding a certain size.

149. FOX, L. and MILLER, J. C. P. 1951 Table-making for large arguments. The exponential integral. *Math. Tab., Wash.* **5**, 163–167.

Gives an example of the use of auxiliary variables for large values of the argument.

Useful instruction on methods of compilation and checking can also be gained from the introductions to published mathematical tables, particularly the series of tables of the British Association, the Royal Society and the National Bureau of Standards.

Interpolation aids

Much practical information is contained in [**167**].

150. FOX, L. 1956 The use and construction of mathematical tables. *Math. Tab. Nat. Phys. Lab.* **1**. London: H.M. Stationery Office.

Describes in detail the various interpolation aids, particularly modified differences and economized polynomials. The analysis of error is very thorough. Also included is a brief survey of the various standard processes for computing mathematical functions. An abridged form of this article is included in a book entitled *On numerical approximation* edited by R. E. Langer, and published by the University of Wisconsin Press, Madison (1959).

151. CLENSHAW, C. W. and OLVER, F. W. J. 1955 The use of economized polynomials in mathematical tables. *Proc. Camb. Phil. Soc.* **51**, 614–628.

Derives expansions in series of central differences for the coefficients of the economized interpolation polynomials, and compares these aids with existing ones. The method given for reducing the rounding error is superseded by that of Chapter 15, § 13. See also [**215**].

152. Woodward, P. M. and Woodward, A. M. 1946 Four-figure tables of the Airy function in the complex plane. *Phil. Mag.* **37**, 236–261.

Discusses the use of modified differences as aids for the interpolation of analytic functions tabulated at points of a square grid in the complex plane.

Subtabulation

The 'end-figure' method is described in [2] and [51].

153. Nautical Almanac Office 1958 *Subtabulation.* London: H.M. Stationery Office.

This comprehensive manual, which is a companion to [167], describes three types of method in order of increasing power and complexity: direct methods, the method of precalculated second differences, and the method of bridging differences. Worked examples and necessary tables are included.

154. Woollett, E. R. 1958 Subtabulation with special reference to a high-speed computer. *Quart. J. Mech.* **11**, 185–195.

TABLES

Tables for desk-machine work

The following books are among the best collected tables of the common functions.

155. *Barlow's tables* 1947 (Edited by L. J. Comrie). Fourth edition (first edition 1814). London: Spon.

These well-known tables give principally squares, cubes, square roots, cube roots and reciprocals of all integers up to 12,500.

156. Comrie, L. J. 1947 *Chambers's four-figure mathematical tables.* Edinburgh: Chambers.

This compact volume includes logarithms, square roots, cube roots, reciprocals, trigonometric, exponential and hyperbolic functions and the error function.

157. Milne-Thomson, L. M. and Comrie, L. J. 1948 *Standard four-figure mathematical tables.* Second edition (first edition 1931). London: Macmillan.

Covers the same ground as the previous reference, but is a larger book because of the use of finer intervals of tabulation. A table of the gamma function is also included.

158. Dale, J. B. 1949 *Five-figure tables of mathematical functions.* Second edition (first edition 1905). London: Arnold.

A compact collection of elementary functions and many higher transcendents, including the gamma and error functions; Bessel functions; and sine, cosine, exponential and elliptic integrals.

159. Comrie, L. J. 1948 and 1949 *Chambers's six-figure mathematical tables* (Two volumes). Edinburgh: Chambers; New York: Van Nostrand.

Covers similar ground to [156] and [157], with of course the extra two-figure accuracy. There are explanations of basic desk-machine practices, similar to those in [167]. Volume I gives logarithmic values, and the more widely-used Volume II gives natural values. An abridged version also exists.

160. Emde, F. 1959 *Tables of elementary functions.* Third edition (first edition 1940). Leipzig: Teubner.

Contains tables of powers, reciprocals and factors, and of trigonometric, exponential and hyperbolic functions with both real and complex arguments. The accuracy is generally 4–5 significant figures. There are many graphs, relief maps, formulae and notes concerning these and other functions, and also a collection of nomograms, formulae and tables for solving quadratic, cubic and quartic equations.

161. Jahnke, E. and Emde, F. 1952 *Tables of higher functions*. Fifth edition (first edition 1909). Leipzig: Teubner.

A well-known and valuable collection of short tables, formulae, graphs and relief maps of the higher transcendental functions of frequent occurrence in numerical work. A new edition, revised by F. Lösch, has just been published.

A large *Handbook of functions* is being compiled by the National Bureau of Standards, Washington. It will contain extensive collections of formulae, tables and graphs for both elementary and higher functions.

162. Fletcher, A., Miller, J. C. P. and Rosenhead, L. 1946 *An index of mathematical tables*. (Second edition in press.) London: Scientific Computing Service.

An excellent index containing a vast amount of information. The second edition covers all tables published before 1955 and major tables in the period 1955–1959. Other information on recent tables is contained in the following Russian index:

163. Lebedev, A. V. and Fedorova, R. M. 1956 Spravochnik po matematicheskim tablitsam. (An index of mathematical tables.) Moscow: Izdatel'stvo Akademii Nauk SSSR,

and its first supplement

164. Burunova, N. M. 1959 Spravochnik po matematicheskim tablitsam. Dopolnenie No. 1. Moscow: Izdatel'stvo Akademii Nauk SSSR.

Tables for automatic work

165. Clenshaw, C. W. 1954 Polynomial approximations to elementary functions. *Math. Tab., Wash.* **8**, 143–147.

Gives 9-decimal values of the coefficients in the Chebyshev expansions for trigonometric, exponential and logarithmic functions, and for the gamma function and the Bessel functions J_0 and J_1. An extension of these tables, giving more functions and an accuracy of generally 20 significant figures, is in preparation for the N.P.L. Mathematical Tables series; see [**214**].

166. Hastings, C. 1957 *Approximations for digital computers*. Second edition (first edition 1955). Princeton University Press.

Includes approximations, usually in rational or explicit polynomial form, for trigonometric, exponential and logarithmic functions, the gamma and error functions, the exponential integral and the complete elliptic integrals. The accuracy of the approximations is 10 significant figures for some functions, but for most of them it is less.

FACTS AND FORMULAE

Many books of tables contain collections of mathematical formulae. The presentation of such information is a primary purpose of the following references.

167. Nautical Almanac Office 1956 *Interpolation and allied tables*. London: H.M. Stationery Office.

This valuable working manual contains tables of interpolation and other coefficients; a large collection of finite-difference formulae for interpolation, differentiation, integration, the solution of ordinary differential equations and estimation of error; and an account of the central-difference method for integrating ordinary differential equations.

168. Dwight, H. B. 1957 *Tables of integrals and other mathematical data*. Third edition (first edition 1934). New York: Macmillan.

A useful small collection of definite and indefinite integrals and series.

169. Adams, E. P. 1939 *Smithsonian mathematical formulae and tables of elliptic functions*. Washington: Smithsonian Institution.

Contains a useful collection of elementary formulae for algebra, trigonometry, geometry, infinite series and some higher transcendental functions.

170. Ryshik, I. M. and Gradstein, I. S. 1957 *Tables of series, products and integrals.* First edition in English. Berlin: Deutscher Verlag der Wissenschaften.

A large collection of formulae for series, indefinite and definite integrals, elementary functions and higher transcendental functions.

Integrals and integral transforms

See [**168**] and [**170**].

171. Meyer zur Capellen, W. 1950 *Integraltafeln.* Berlin: Springer.

Gives indefinite integrals of elementary functions.

172. Grobner, W. and Hofreiter, N. 1957 and 1958 *Integraltafeln* (Two Parts). Second edition (first edition 1949). Vienna: Springer.

Part I (1957) and Part II (1958) give respectively indefinite and definite integrals of elementary functions.

173. Erdélyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F. G. 1954 *Tables of integral transforms* (Two volumes). New York and London: McGraw-Hill.

A vast collection of integral transforms of the higher transcendental functions.

174. Byrd, P. F. and Friedman, M. D. 1954 *Handbook of elliptic integrals for engineers and physicists.* Berlin: Springer.

175. Kamke, E. *Differentialgleichungen, Lösungsmethoden und Lösungen.* Part I, 1956 (fifth edition). Leipzig: Becker and Erler. Part II, 1959 (fourth edition). Leipzig: Akademische Verlagsgesellschaft.

Gives analytic solutions of various differential equations. Part I deals with ordinary and Part II with partial differential equations.

CURVE-FITTING AND SMOOTHING

This subject is not easy and has pitfalls for the unwary. Good introductions to the practical side are contained in [**2**] and [**6**], both of which illustrate some of the dangers in the process. Sound theoretical accounts of least-square approximations and the use of orthogonal polynomials are contained in [**4**] and [**1**]. See also [**11**] and [**44**].

176. Fisher, R. A. and Yates, F. 1957 *Statistical tables for biological, agricultural and medical research.* Fifth edition (first edition 1938). Edinburgh: Oliver and Boyd.

Table XXIII gives values of orthogonal polynomials up to degree 5 to facilitate the fitting of equispaced data at up to 75 points.

177. DeLury, D. B. 1950 *Values and integrals of the orthogonal polynomials up to* $n = 26$. Toronto University Press.

Extends the tables mentioned in the preceding reference, when the number of points does not exceed 26, by giving the values of all the orthogonal polynomials.

178. Hayes, J. G. and Vickers, T. 1951 The fitting of polynomials to unequally-spaced data. *Phil. Mag.* **42**, 1387–1400.

Describes good desk-machine procedures for fitting unequally-spaced data.

179. Forsythe, G. E. 1957 Generation and use of orthogonal polynomials for data-fitting with a digital computer. *J. Soc. Indust. Appl. Math.* **5**, 74–88.

180. Ascher, M. and Forsythe, G. E. 1958 SWAC experiments on the use of orthogonal polynomials for data fitting. *J. Ass. Comp. Mach.* **5**, 9–21.

These two papers examine the problem of fitting unequally-spaced data from the standpoint of automatic computation.

181. Clenshaw, C. W. 1960 Curve fitting with a digital computer. *Computer J.* **2**, 170–173.

Describes a refinement of the method given in the two preceding references which makes lower demands on the store and achieves a more concise form of output.

HARMONIC ANALYSIS

See [1], [2], [4], [11] and, especially, [6].

182. Brunt, D. 1931 *The combination of observations*. Second edition (first edition 1923). Cambridge University Press.

183. Pollak, L. W. *Geophysical publications*. 1947 and 1949. Volume 1 (1947). Harmonic analysis and synthesis schedules for 3 to 100 equidistant values of empiric functions. Volume 2 (1949). All term guide for harmonic analysis and synthesis. Dublin: Stationery Office.

184. Danielson, G. C. and Lanczos, C. 1942 Some improvements in practical Fourier analysis and their application to X-ray scattering from liquids. *J. Franklin Inst.* **233**, 365–380 and 435–452.

INTEGRAL EQUATIONS

Numerical treatments are included in [1], [3], [5].

185. Lovitt, W. V. 1950 *Linear integral equations*. Reprint (first edition 1924). New York: Dover.

An elementary introduction to analytical theory.

186. Tricomi, F. G. 1957 *Integral equations*. New York and London: Interscience.

A compact account of the analytical theory.

187. Fox, L. and Goodwin, E. T. 1953 The numerical solution of non-singular linear integral equations. *Phil. Trans.* A, **245**, 501–534.

A comprehensive account of the solution of equations of Fredholm and Volterra types by finite-difference methods.

188. Young, A. 1954 Approximate product-integration. *Proc. Roy. Soc.* A, **224**, 552–561.

This is a preliminary to the following paper.

189. Young, A. 1954 The application of product-integration to the numerical solution of integral equations. *Proc. Roy. Soc.* A, **224**, 561–573.

Applicable to certain types of singular or near-singular equations.

MISCELLANEOUS

190. Stegun, I. A. and Abramowitz, M. 1956 Pitfalls in computation. *J. Soc. Indust. Appl. Math.* **4**, 207–219.

191. Forsythe, G. E. 1958 Singularity and near singularity in numerical analysis. *Amer. Math. Mon.* **65**, 229–240.

COMPUTING MACHINES

Desk machines

Brief introductions are given in [1] and [2]. Fuller treatments on use are given in the following two references and [197] below.

192. Sabielny, H. 1939 *Modern machine calculation*. London: Scientific Computing Service.

193. Varner, W. W. 1957 *Computing with desk calculators*. New York: Rinehart.

Punched-card machines

194. Smith, J. S. 1960 *Punched cards*. London: Macdonald and Evans.

Contains a readable account of the various types of punched-card machines, and discusses their applications, particularly to accountancy.

195. CASEY, R. S., PERRY, J. W., BERRY, M. M. and KENT, A. (Editors) 1959 *Punched cards: their applications to science and industry*. Second edition (first edition 1951). New York: Reinhold; London: Chapman and Hall.

Contains an extensive bibliography.

See also [**197**].

Automatic digital computers

See [**8**], [**209**], [**210**], and [**211**].

196. GRABBE, E. M., RAMO, S. and WOOLDRIDGE, D. E. (Editors) 1959 *Handbook of automation, computation, and control*. Volume 2: *Computers and data processing*. New York and London: John Wiley.

This reference work of some 1000 pages also includes chapters on analogue computers.

197. MONTGOMERIE, G. A. 1956 *Digital calculating machines*. Glasgow: Blackie.

Provides a short introduction; desk machines and punched-card machines are also treated.

198. HOLLINGDALE, S. H. 1959 *High speed computing: methods and applications*. London: English Universities Press.

An introduction aimed at the general scientific reader.

199. WILKES, M. V. 1956 *Automatic digital computers*. London: Methuen.

Concentrates on logical design and programming.

Analogue computers

See [**196**].

200. SOROKA, W. W. 1954 *Analog methods in computation and simulation*. New York and London: McGraw-Hill.

A good introduction, which describes all types of analogue machines and discusses their applications.

201. HARTREE, D. R. 1949 *Calculating instruments and machines*. Urbana: University of Illinois Press.

The first part describes the application of the differential analyser and similar machines. The second part similarly treats automatic digital computers, but it is now somewhat out of date.

202. JOHNSON, C. L. 1956 *Analog computer techniques*. New York and London: McGraw-Hill.

Devoted to the electronic differential analyser.

203. KARPLUS, W. J. 1958 *Analog simulation*. New York and London: McGraw-Hill.

Devoted to machines for solving partial differential equations. Applications are treated in considerable detail.

NOMOGRAPHY

204. BRODETSKY, S. 1938 *A first course in nomography*. London: Bell.

An elementary account of the construction of alignment nomograms for equations containing up to four variables.

205. ALLCOCK, H. J., JONES, J. R. and MICHEL, J. G. L. 1950 *The nomogram*. Fourth edition (first edition 1932). London: Pitman.

A readable book containing all the computor really requires about the subject.

206. D'OCAGNE, M. 1921 *Traité de nomographie*. Paris: Gauthier-Villars.

This is an outstanding account of the theory of nomography.

207. Dulaey, M. 1951 *Construction des abaques*. Paris: Gauthier-Villars.

208. Meyer zur Capellen, W. 1953 *Leitfaden der Nomographie*. Berlin: Springer.

A good summary, with bibliographies, of modern Russian developments in the subject, including the empirical construction of nomograms, is contained in *Vychislitel'naya Matematika*, No. 4 (1959).

JOURNALS AND REVIEWS

The volume of literature on numerical analysis has increased very rapidly in the last two decades, and papers appear in many periodicals. Journals which are largely devoted to the subject include:

Journal of the Association for Computing Machinery (American),
The Computer Journal (British),
Mathematics of Computation (formerly *Mathematical Tables and other Aids to Computation*) (American),
Numerische Mathematik (German),
Journal of the Society for Industrial and Applied Mathematics (American),
Chiffres (French),
Vychislitel'naya Matematika (Russian).

Reviews of papers on numerical analysis appear in

Mathematical Reviews,
Applied Mechanics Reviews,
Computing Reviews (at present incorporated in the *Communications of the Association for Computing Machinery*),
Referativnyi Zhurnal: Matematika (Russian),

Mention may also be made of the

International Journal of Abstracts: Statistical Theory and Method, which gives a complete coverage in the field of statistical theory and new contributions to statistical method as published after October 1st, 1958.

ITEMS ADDED IN PROOF

209. Lance, G. N. 1960 *Numerical methods for high speed computers*. London: Iliffe.

Gives an account of methods of numerical analysis which are suitable for automatic work. Subjects include evaluation of functions, matrix algebra, ordinary and partial differential equations, zeros of polynomials, continued fractions and quadrature.

210. Ralston, A. and Wilf, H. S. (Editors) 1960 *Mathematical methods for digital computers*. New York and London: John Wiley.

Gives some selected methods, with flow diagrams. The subjects treated are similar to those of the preceding reference.

211. Alt, F. L. (Editor) 1960 *Advances in computers*. Volume 1. New York and London: Academic Press.

Surveys recent progress in business applications, weather prediction, language translation, games applications and word recognition. Further volumes on these and other applications, including numerical analysis, are in preparation.

212. Faddeeva, V. N. 1959 *Computational methods of linear algebra*. (Translated by C. D. Benster.) New York: Dover; London: Constable.

The first part of this book is a useful summary of the parts of matrix theory which are important in numerical analysis. The second part provides an account of some of the more important methods, direct and iterative, for solving linear equations and inverting matrices. The third and final part treats methods for computing latent roots and vectors, but is less valuable than the other parts because of rapid advances which have been made in this field since the original Russian edition was written.

213. Wilkinson, J. H. 1960 Error analysis of floating-point computation. *Numerische Math.* **2**, 319–340.

Gives the fundamental inequalities satisfied by the rounding errors in the basic arithmetical operations. Applies the results to the analysis of a number of related techniques for computing latent roots.

214. Clenshaw, C. W. Chebyshev series for mathematical functions. *Math. Tab. Nat. Phys. Lab.* **5**, London: H.M. Stationery Office. (*In press.*)

Gives 20-decimal values of the coefficients in the Chebyshev expansions for the trigonometric, exponential, logarithmic, gamma and error functions, the exponential integral, and Bessel functions J, Y, I, K of orders 0 and 1. The Introduction gives a comprehensive account of the use of Chebyshev series in numerical analysis.

215. Olver, F. W. J. Tables for Bessel functions of moderate or large orders. *Math. Tab. Nat. Phys. Lab.* **6**, London: H.M. Stationery Office. (*In press.*)

Provides interpolation aids based on economized polynomials (Chapter 15, §§ 12–15).

216. Wilkinson, J. H. *The algebraic eigenvalue problem.* Oxford University Press. (*In press.*)

A critical assessment of methods from the standpoint of automatic computation, with emphasis on numerical stability.

217. Forsythe, G. E. and Wasow, W. R. 1960 *Finite-difference methods for partial differential equations.* New York and London: John Wiley.

A comprehensive account of modern finite-difference procedures, with special emphasis on automatic computation. All three kinds of partial differential equation are considered, though naturally the most extensive treatment is that of elliptic equations.

INDEX

Acceleration of convergence, 123, 155–156, *see also* Aitken's δ^2-process, Extrapolation, Summation of series
Adams–Bashforth method, 83
Aitken's δ^2-process, 40, 123, 132, 149, 156
Algebraic equations, *see* Linear algebraic equations, Zeros of polynomials
Alternating-direction method, 155
Analogue computers, 163
Analytic solution of differential equations: ordinary, 161; partial, 153, 161
Asymptotic expansions, asymptotic series, 125, 135, 158
Automatic computers, 13–14, 163; tables for, 144, 160

Back-substitution, 5, 15; error analysis of, 46–48
Bairstow's process, 57–61
Bessel functions, computation by recurrence, 128
Bessel's interpolation formula, 67, 139
'Best' or 'best fit' polynomial approximation, 72–74
Biharmonic equation, 120–121, 154–155
Bisection, method of, 31–32
Boundary-value problems in ordinary differential equations, 81, 93–100, 152
——, methods for solving: Chebyshev series, 99; direct finite-difference, 93–97; initial-value techniques, 98–99; relaxation, 152
Building-up errors: in solution of ordinary differential equations, 90, 152; in the use of recurrence relations, 128–129

Central-difference methods for ordinary differential equations, 83–85, 151
Characteristic equation, characteristic polynomial, 23
Characteristic strip, 103
Characteristics: definition of, 103; method of, 105–108, 153
Chebyshev polynomials, 71–75; explicit expressions for, 75; properties of, 74–75
Chebyshev series, 71–79, 150–151; calculation of coefficients in, 75–76; differentiation of, 79; evaluation of, 76–77; integration of, 78; methods for solving ordinary differential equations, 88–90, 99, 100; summation of, 76–77
Computing machines, 162–163
Continued fractions, 127–128, 156
Converging factor, 156
Crank–Nicolson method, 116, 153
Crout's method, 8
Cubic equations, 54, 149
Curve-fitting, 161

Deferred approach to the limit, 122–123
Deferred-correction methods, 85–87
Desk machines, 162
Determinant, 3
Difference correction, 85, 122
Difference-correction techniques for solving differential equations: ordinary, 85–87, 154; partial, 116, 118, 154
Differences, *see* Finite differences
Differential equations, *see* Ordinary, Partial
Differentiation, numerical: by Chebyshev series, 79; finite-difference formulae for, 67–68
Doolittle's method, 7, 11
Double integrals, 158

Economized polynomials, 141
Eigenvalues: of matrices, *see* Latent roots; of ordinary differential equations, 100; of partial differential equations, 121
Eigenvectors, *see* Latent vectors
Elimination, 5–6; compact methods for, 7; *see also* Gaussian elimination
Elliptic partial differential equations, 103, 111, 117–121, 154–155; eigenvalue problems in, 121; stability of methods for, 152–153; treatment of singularities, 155; with curved boundaries, 119, 154; with free boundaries, 121
Error analysis: of linear algebraic processes, 41–52, 148–149; of back substitution, 46–48; of floating-point computations, 50–51; of Gaussian elimination, 42–48; of ill-conditioned equations, 51; of interpolation, 141–142; of polynomial equations, 59; of triangular decomposition, 48–50
Errors: building-up, 90, 128, 152; detection by differences, 63, 138, 158
Euler–Maclaurin formula, 127
Euler's transformation of series, 124, 132, 133, 136; van Wijngaarden's modified procedure for, 125

Everett's interpolation formula, 66, 132, 139

Extrapolation: exponential, 123, 132, *see also* Aitken's δ^2-process; h^k-, 122–123

Finite-difference formulae: for interpolation, 65–67; for numerical differentiation, 67–68; for numerical integration, 68–70; for ordinary differential equations, 83–87, 93–97; for partial differential equations, 113–121

Finite-difference operators, 64–70

Finite differences, 62–63; as interpolation aids, 139; calculus of, 64–70, 149–150; checking by, 63; divergence of, 70; effect of errors on, 63, 158; notation for, 63; symbolic relations between, 64

Fixed-point arithmetic, 14–15, 20

Floating-point arithmetic, 14, 20, 50–51

Fourier series, 73–74, 125, 162

Gaussian elimination, 5–6; error analysis of, 42–48; variants of, 14, 17; with interchanges, 15–17

Gaussian quadrature, 131–132, 134, 157

Gauss–Seidel method, 36–39

Gillies' method for reducing rounding errors, 141

Givens' method, 30–31, 147

Gregory's integration formula, 69

h^2-extrapolation, 122–123

Harmonic analysis, 162

Hessenberg matrix, 147, 148

Householder's method, 33

Hyperbolic partial differential equations, 101–111, 153; for compressible flow, 109–111; discontinuity of solutions of, 103; simultaneous, 110; stability of methods for, 152–153; for vibrating string, 104–106

Ill-conditioning: of matrices and linear equations, 5, 7, 11, 19–21; error analysis of, 51; of polynomial equations, 59

Indirect methods for linear algebraic equations, 34–40, 155; Gauss–Seidel, 36–39; Jacobi's, 35–38; Liebmann's, 36, 39; relaxation, 34; simultaneous displacements, 35; successive displacements, 36; successive overrelaxation, 39–40

Initial-value problems in ordinary differential equations, 80–92, *see also* Ordinary differential equations

Integral equations, 162

Integral transforms, 161

Integrals, 130–136, 157–158; analytical evaluation of, 135, 161; double, 158; evaluation by Chebyshev series, 78

Integration, formulae for: finite-difference, 68–70; Gauss-type, 131–132, 134, 157; Gregory's, 69; Newton–Cotes, 70, 134; Simpson's, 69–70, 91, 130; trapezoidal, 68, 123, 131

Integration of differential equations, *see* Ordinary differential equations, Partial differential equations

Interchanges: in Gaussian elimination, 15–17; in triangular decomposition, 18–19, 48

Interpolating polynomial, 65

Interpolation: aids, 139–142, 158–159; auxiliary variables for, 142–143; by economized polynomials, 141; by finite differences, 139; in two variables, 159; inverse, 67; near singularities, 142; successive linear, 53, 56; tables for, 160; using reduced derivatives, 140; *see also* Interpolation formulae, Subtabulation

Interpolation formulae, 65–67, 139–142; Bessel's, 67, 139; Everett's, 66, 132, 139; Lagrange's, 142; Newton's, 65

Inverse interpolation, 67

Iterative processes: acceleration of, *see* Acceleration of convergence; classification of, 155; for latent roots and vectors, 24–26; for linear algebraic equations, *see* Indirect methods; for zeros of polynomials, 53–61; order of convergence of, 56, 122

Jacobi's method: for latent roots and vectors, 29–30; for linear algebraic equations, 35–38

Lagrange's interpolation formula, 142

Laplace's equation, 117–120

Latent roots and vectors, 22–33, 147–148; linear independence of, 23, 37; subdominant, 26–29

——, methods for evaluating: bisection, 31–32; Givens', 30–31, 147; Householder's, 33; iteration, 24–26; Jacobi's, 29–30; LR-transformation, 148; minimized iterations, 148; QD-algorithm, 157

Least-squares approximations, 161

Liebmann's method, 36, 39

Limits, evaluation of, 122–129, 155–157

Linear algebraic equations, 1–21, 34–40, 147; automatic solution of, 13–21; error analysis of, 41–52, 148–149; ill-conditioning of, 5, 7, 11, 19–21

——, methods for: Crout's, 8; Doolittle's, 7, 11; triangular resolution or decomposition, 8–12, 18–19, 48–50; *see also* Gaussian elimination, Indirect methods

Linear independence, 4, 23, 37, 138

LR-transformation, 148

Machine tables, 77, 144, 160
Mathematical tables: interpolation aids for, 139–142; of elementary functions, 159–160; preparation and printing of, 137, 138
Matrix, 1; addition of, 2; band or co-diagonal, 31, 94, 97, 99; characteristic equation of, 23; Hessenberg, 147, 148; identity, 2; inversion of, 2–5, 11–12, 17, 147; mathematical theory of, 146–147; methods for elliptic partial differential equations, 118; multiplication of, 1–3, 9; norms, 148; orthogonal, 29; positive definite, 9, 23; property 'A', 38–39; quadratic form associated with, 23; reciprocal, 2; rotation of, 30; similarity transformation of, 148; singular, 4; symmetric, 3; trace of, 30; transposed, 3; triangular, 8; triangular resolution of, 8–12, 18–19, 48–50; tri- or triple-diagonal, 31, 94, 97–98; unit 2; *see also* Latent roots and vectors
Minor, 3–4; leading principal, 9
Modified differences, 139–140
Modulus function, 133, 143

Nested multiplication, 53, 76
Newton–Cotes integration formulae, 70, 134
Newton's interpolation formulae, 65
Newton's rule, 55–57, 60
Nomography, 163–164
Numerically satisfactory solutions of differential equations, 137–138

Operators, in finite-difference theory, 64–70
Order of convergence of approximate formulae, 56, 122
Ordinary differential equations, 80–100, 151–152; analytic solution of, 161; building-up errors in, 90, 152; choice of solutions of, 137–138; eigenvalue problems in, 100, 152
——, methods for solving: Adams–Bashforth, 83; central-difference, 83–85, 151; Chebyshev series, 88–90, 99–100; deferred-correction, 85–87; predictor-corrector, 82–83, 151; Runge–Kutta, 87–88, 151–152; Taylor-series, 81–82, 151; de Vogelaere's, 88; stability of, 90–92, 152; *see also* Boundary-value problems, methods for solving
Orthogonal polynomials, 73
Overrelaxation parameter, 39–40

Parabolic partial differential equations, 103, 112–117, 153–154; analytic solution of, 153; non-linear, 116; treatment of singularities, 116; with moving boundaries, 154
——, methods for solving: Crank–Nicolson, 116, 153; explicit, 115; implicit, 116; Richardson's, 115; Runge–Kutta, 114–115; stability of, 115–116, 154
Partial differential equations: analytic solution of, 153, 161; boundary conditions for, 111, 112; classification of, 101–103; discontinuity of solutions of, 103; quasi-linear, 101, 109; *see also* Elliptic, Hyperbolic and Parabolic partial differential equations
Phase function, 133, 143
Pivotal condensation, 5, *see also* Gaussian elimination
Polynomials: division by linear factor, 53; division by quadratic factor, 57; economized, 141; evaluation of, 53, 57; evaluation of zeros of, *see* Zeros of polynomials; of best approximation or fit, 72–74; *see also* Chebyshev polynomials
Predictor-corrector methods, 82–83, 151
Proof-reading, 138
Property 'A', 38–39
Punched-card machines, 162–163

Quadratic convergence, 55
Quadrature, *see* Integration, formulae for
Quartic equations, 54–55, 149
Quasi-linear partial differential equations, 101, 109
Quotient-difference ('QD') algorithm, 156–157

Recurrence relations, 128–129, 135
Reduced derivatives, 82, 140
Reducing cubic, 54
Relaxation methods: for linear algebraic equations, 34; for ordinary differential equations, 100; for partial differential equations, 118, 154–155; *see also* Successive overrelaxation
Richardson's method: for h^2-extrapolation, 122–123; for parabolic equations, 115
Rounding errors: in an interpolate, 141; in the solution of linear algebraic equations, 14, 19, 41–51, 148–149, polynomial equations, 59–60; *see also* Error analysis, Errors
Runge–Kutta method, 87–88, 91–92, 151–152; used to solve parabolic partial differential equations, 114–115

Scalar product, 3; exact accumulation of, 14, 15, 19–21
Simpson's rule, 69–70, 91, 130
Singular integrals, 134, 150, 158
Singularities: in interpolation, 142–143; in numerical analysis generally, 162; in partial differential equations, 103, 116–117, 154, 155
Smoothing, 161
Stability of methods: for ordinary differential equations, 90–92, 152; for partial differential equations, 115–116, 154; for zeros of polynomials, 59–60

Subtabulation, 138, 159
Successive overrelaxation, 39–40
Summation of series, 124–127, 155–156

Table-making, 137–139, 158
Tables, *see* Machine tables, Mathematical tables
Taylor series, 71–72
Taylor-series method, 81–82, 151
Taylor's theorem, operational form of, 65
Throw-back, 139
Trapezoidal rule, 68, 123, 131
Triangular decomposition, 8–12; error analysis of, 48–50; with interchanges, 18–19

Vector, 1; biorthogonal, 24; displacement, 35; error, 35–39; normalized, 22; principal, 37, 39; residual, 49
de Vogelaere's method, 88

Weight factors for numerical integration, 131, 157
van Wijngaarden's modified procedure for Euler's transformation, 125
van Wijngaarden's transformation of series, 126, 132

Zeros of polynomials, 53–61, 149; effect of changes in the coefficients on, 59; ill-conditioning of, 59; multiple, 59; of 3rd or 4th degree, 54, 149
——, methods for evaluating: Bairstow's, 57–61; Bernoulli's, 149; choice of, 54, 60–61; convergence of, 55–57; Muller's, 149; Newton's, 55–57, 60; QD algorithm, 156–157; root-squaring, 61, 149; stability of, 60; successive linear interpolation, 53, 56